FOUR DIFFICULT LESSONS 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 difficult lessons: 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

2. METHODOLOGY

Thirteen chord qualities, given in Table **??**, in all twelve pitch classes and one no-chord class are considered for a total of 157 chord classes. Having all four datasets at hand, these collections are merged into the largest collection of chord transcriptions used to date, totaling 1235 tracks. Given that the collections were curated in isolation of each other, it is a necessary first step to identify and remove duplicates to avoid data contamination during cross validation. To these ends, each recording is checked against the EchoNest Analyze API [1] and associated with its track and song identifiers, corresponding to the recording and work, respectively. Though multiple track IDs will map to the same song ID, uniqueness is defined at the level of a song to ensure duplicates are removed. This identifies 18 redundant songs, and all but one is dropped for each collision from the total collection, resulting in a final count of 1217 unique tracks.

2.1 Automatic Systems

For algorithmic parity, two systems are considered here having adopted the same output chord prediction space and trained over the same data splits.

2.1.1 K-stream GMM-HMM with Multiband Chroma

Following the lineage of automatic chord estimation systems, one system considered is that of [**?**]. A multiband chroma representation is computed from beat-synchronous audio analysis, producing four parallel chroma features. Each is fit to a separate Gaussian Mixture Model (GMM) by rotating all chroma vectors and chord labels to C. During inference, four separate observation likelihoods over all chord classes are obtained by circularly rotating the feature vector the GMM. These four posteriors are then decoded jointly, using a k-stream HMM, resulting in a beat-aligned chord sequence. In addition to being one of the highest performing systems at a recent iteration of MIReX, a software implementation was obtained, thereby enabling experimental consistency between partitions of the training data.

2.1.2 Deep Convolutional Neural Network

Acknowledging both the limited representational power of GMMs and a similar trend that occurred in automatic speech recognition, a deep convolutional network is also considered [**?**]. Time-frequency patches of local contrast normalized constant-Q spectra, on the order of one second, are transformed by a fully-convolutional network. Finding inspiration in the root-invariance strategy of GMM training, explicit weight-tying is achieved across roots such that all qualities develop the same internal representations, allowing the model to generalize to chords unseen during.

3. PERFORMANCE ANALYSIS

The proceedings will be printed on portrait A4-size paper (21.0cm x 29.7cm). All material on each page should fit within a rectangle of 17.2cm x 25.2cm, centered on the

[1] http://developer.echonest.com/docs/v4

page, beginning 2.0cm from the top of the page and ending with 2.5cm from the bottom. The left and right margins should be 1.9cm. The text should be in two 8.2cm columns with a 0.8cm gutter. All text must be in a two-column format. Text must be fully justified.

4. OBSERVATIONS

Based on this analysis, a handful of instances are pulled out of the data and inspected more closely.

4.1 Assumptions on Tonality

Not all music is well described by chords. Because you *can* annotate chords in a song, doesn't mean you should. Revolution 9, Beastie Boys, Fugees. These are not well described by chords, and introduce noise in the evaluation process.

Alternatively, systems are sensitive to tuning. This is another easy way to pick up goose eggs during evaluation. While the outputs might be spot on and even useful to a human, sensitivity to absolute chord spelling fails to This motivates harmonic, Roman numeral analysis as a slightly different formulation of the task; decouple tuning from function. While not all chord datasets contain this information, Billboard and Isophonics do.

4.2 The Significance of Chord Representations

Comparing chords in label space is bonkers. Mappings or resolutions effectively quantize chords to a one-hot encoding. These can be thought of as mapping chords to bit vectors and then testing for equivalence. This is the wrong representation.

One, chords are inherently hierarchical, and this approach to resolution discards these relationships. Flat classification problems —those in which different classes are conceptually independent— are built on the assumption of mutually exclusive relationships. In other words, assignment to one class precludes the valid assignment to any other classes considered. For example, "cat" and "dog" are mutually exclusive classes of "animal", but "cat" and "mammal" are not. Returning to chords, `C:dim7` and `C:maj` are clearly mutually exclusive classes, but it is difficult to say the same of `C:maj7` and `C:maj`, as the former *contains* the latter.

Two, the flexibility of the standard Harte syntax can be abused for ambiguous chords, and it isn't clear what to do with these labels.

Occurrence of bizarre chord spellings in the data: Should these even exist?

Chords with bass intervals other than the root should be discarded from chord mapping strategies. Otherwise, this will only introduce noise.

4.3 The Natural Conflict between Recognition and Transcription

The former is literal, the latter is anything but. In a recognition problem, silence is always no-chord, because nothing is playing. Transcription, on the other hand, is attempting to assign labels to regions, and is closer to segmentation than classic approaches to chord estimation. It is easy to find instances of both in the data.

Table 1. Various real chord transcriptions for "With or Without You" by U2, comparing the reference annotation 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.

Ver.	Chord Sequence				Score	Ra
Ref.	`D:maj`	`D:maj/5`	`D:maj6/6`	`D:maj(4)/4`	—	
1	`D:maj`	`A:maj`	`B:min`	`G:maj`	4/5	1
2	`D:5`	`A:sus4`	`B:min7`	`G:maj`	5/5	
3*	`D:maj`	`A:maj`	`B:min`	`G:maj`	4/5	
4*	`D:maj`	`A:maj`	`B:min`	`G:maj7`	4/5	
5*	`D:maj`	`A:maj`	`B:min`	`G:maj`	5/5	2
6	`D:5`	`A:5`	`D:5/B`	`G:5`	5/5	

Unfortunately, reference datasets contain annotations of both styles, and sometimes internally to the same annotation. Over-specified chords are mostly indicative of a recognition problem, and finds an ambiguous middle ground between harmonic analysis and pitch recognition. Alternatively, there are plenty of instances in which annotators make transcriptions decisions. Nirvana – all apologies.

4.4 "There is No Spoon," or the

Known issue, Ni and company. Not enough attention is being drawn to this observation. How can we objectively quantify a subjective task?

Examples in the data

5. FIRST LEVEL HEADINGS

First level headings are in Times 10pt bold, centered with 1 line of space above the section head, and 1/2 space below it. For a section header immediately followed by a subsection header, the space should be merged.

5.1 Second Level Headings

Second level headings are in Times 10pt bold, flush left, with 1 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

5.1.1 Third and Further Level Headings

Third level headings are in Times 10pt italic, flush left, with 1/2 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

Using more than three levels of headings is highly discouraged.

String value	Numeric value
Hello ISMIR	2015

Table 2. Table captions should be placed below the table.

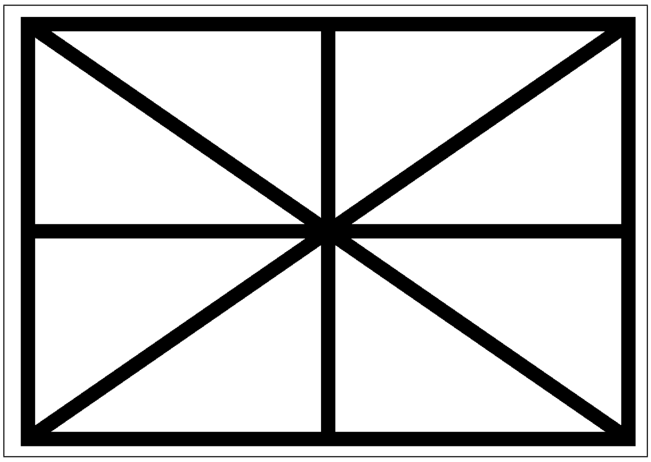

Figure 1. Figure captions should be placed below the figure.

6. FOOTNOTES AND FIGURES

6.1 Footnotes

Indicate footnotes with a number in the text.[2] Use 8pt type for footnotes. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a 0.5pt horizontal rule.

6.2 Figures, Tables and Captions

All artwork must be centered, neat, clean, and legible. All lines should be very dark for purposes of reproduction and art work should not be hand-drawn. The proceedings are not in color, and therefore all figures must make sense in black-and-white form. Figure and table numbers and captions always appear below the figure. Leave 1 line space between the figure or table and the caption. Each figure or table is numbered consecutively. Captions should be Times 10pt. Place tables/figures in text as close to the reference as possible. References to tables and figures should be capitalized, for example: see Figure 1 and Table 2. Figures and tables may extend across both columns to a maximum width of 17.2cm.

7. CONCLUSIONS

In this work, the application of deep learning to large-vocabulary ACE is thoroughly explored, advancing the state of the art using standard evaluation methods. Arguably of more importance, both the behavior of the resulting systems and the data used for development are explored in rigorous detail. Our results show that the state of the art may have truly hit a glass ceiling, due to the conventional assumption that "ground truth" data can be obtained for what is, at times, an unavoidably subjective task. This challenge is further compounded by approaches to prediction and evaluation, which attempt to perform flat classification of a hierarchically structured chord taxonomy. Thus, while there certainly remains room for improvement, error analysis indicates that the vast majority of error in modern chord recognition systems is a result of invalid assumptions baked into the very question being asked.

Notably, four issues with current chord estimation methodology have been identified in this work. One, it seems necessary that computational models, and especially those that estimate a large number of chord types, embrace structured outputs; one-of-K class encoding schemes introuduce unnecessary complexity between what are naturally hierarchical relationships. Two, there is value in distinguish between the two tasks at hand, being chord recognition —I am playing this *exact* chord shape on guitar— and chord transcription —finding the best chord label to describe this harmonically homogenous region of music— and how this intent is conveyed to the authors of reference annotations. Three, as championed by [**?**], chord transcription would certainly seem to benefit from explicit segmentation, rather than letting such boundaries between regions of harmonic stability result implicitly from post-filtering algorithms, i.e. Viterbi. Lastly, the all-too-often subjective nature of chord labeling needs to be acknowledged in the process of curating reference data, and the human labeling task should average or combine multiple perspectives rather than attempt to yield canonical "expert" references.

[2] This is a footnote.