BEYOND CHORD VOCABULARIES: EXPLOITING PITCH-RELATIONSHIPS IN A CHORD ESTIMATION METRIC

Johanna Devaney
Brooklyn College and the Graduate Center, CUNY
johanna.devaney@brooklyn.cuny.edu

ABSTRACT

Chord estimation metrics treat chord labels as independent of one another. This fails to represent the pitch relationships between the chords in a meaningful way, resulting in evaluations that must make compromises with complex chord vocabularies and that often require time-consuming qualitative analyses to determine details about how a chord estimation algorithm performs. This paper presents an accuracy metric for chord estimation that compares the pitch content of the estimated chords against the ground truth that captures both the correct notes that are estimated and additional notes that are inserted into the estimate. This is not a stand-alone evaluation protocol but rather a metric that can be integrated as a weighting into existing evaluation approaches.

1. INTRODUCTION

Chord estimation has a long history at ISMIR [1], yet even current approaches still have not exceeded 90% accuracy on simple chord label prediction tasks. Part of the challenge of chord label classification is the large number of label permutations [2] and annotator subjectivity for rare chords, as well as idiosyncratic annotation styles [3] and disagreements between expert annotators [4–6]. Another part is that the evaluation metrics typically employed treat chord labels as independent, rather than as collections, or sets, of pitches. Thus incorrect chord labels are treated the same in these metrics, regardless of whether or not they have any common pitch content with the ground truth. This paper presents an accuracy metric that can be integrated as a weighting mechanism with existing chord estimation evaluation approaches. The benefit of this approach is that it captures pitch relationships between the predicted chords and provides a more nuanced evaluation than treating the chord labels as independent from one another.

1.	Chord root note only
2.	Major and minor: N, maj, min
3.	Seventh chords: N, maj, min, maj7, min7, 7
4.	Major and minor with inversions: N, maj, min, maj/3, min/b3, maj/5, min/5
5.	Seventh chords with inversions: N, maj, min, maj7, min7, 7, maj/3, min/b3, maj7/3, min7/b3, 7/3, maj/5, min/5, maj7/5, min7/5, 7/5, maj7/7, min7/b7, 7/b7

Table 1. List of the chord vocabulary classes, based on [14], that are used for evaluation in the current MIREX Audio Chord Estimation task.

2. BACKGROUND

Chord-labels are a common harmonic representation in both the symbolic and audio domains. The use of chord labels in symbolic music comes out of a long tradition of Roman numeral-focused pedagogy, particularly in North America (e.g., [7, 8]). Models developed in music theory and cognition, such as Krumhansl's [9] and Lerdahl's [10] work, have informed the development of computational distance metrics for chords (e.g., [11–13]). In contrast, much of the harmonic analysis work in the field of music information retrieval has focused on chord recognition because of its simple mapping to a classification problem. Current chord estimation metrics, as exemplified by the ones currently in use MiREX (shown in Table 1), focus on the prediction of chord labels, with varying degrees of simplifications applied to account for chord vocabulary size [14].

One problem with the formulation of chord estimation as a simple classification task with independent labels is that overlapping pitch content between chords is ignored. The chord labels themselves (e.g., C d- e- F G a bô or I ii iii IV V vi viiô) do not themselves provide information about the relationship between the chords. However, their pitch content can be extracted and compared in order to reveal that, for example, the chords d- (ii in C Major) and F (IV in C Major) are musically closer, due to their shared pitch content, than F (IV) and G (V), even though F and G are closer to one another in the label set. The broader issue of conceiving harmonic analysis in terms of labels rather than pitch content has previously been discussed in [15] and [16].

Recent work, such as [17], has attempted to leverage the

note name	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
pitch class	0	1	2	3	4	5	6	7	8	9	10	11
C (I)	o	-	-	-	o	-	-	o	-	-	-	-
d (ii)	-	-	o	-	-	o	-	-	-	o	-	-
e (iii)	-	-	-	-	o	-	-	o	-	-	-	o
F (IV)	o	-	-	-	-	o	-	-	-	o	-	-
G(V)	-	-	o	-	-	-	-	o	-	-	-	o
a (vi)	o	-	-	-	o	-	-	-	-	o	-	-
b^o (vii^o)	-	-	o	-	-	o	-	-	-	-	-	o

Table 2. Summary of the pitch content in the diatonic triads in the C Major scale. Note names are listed in the top row and the numbers in the second row represent the 12 pitch classes. In the lower part of the table, a o indicates the presence of a pitch class in a triad. In this representation, it is clear that d (ii) and F (IV) are musically closer to each other than F (IV) and G (V) because they share more pitch classes in common, i.e., F and A.

overlap in pitch content between varying types of chords. Although this has been limited to defining chord alphabets amongst different chord qualities or types (such as major, minor, or diminished) for a single chord rather than between different chords.

3. PROPOSED ACCURACY METRIC

In any of the evaluation classes shown in Table 1, if an F major chord (IV in C major) is misestimated as a d minor chord (ii), it would be considered equally incorrect as if it were misestimated as a G major chord (V). This fails to capture the fact there are two of the three notes in common between d minor and F major chords, while no notes are in common between the G major and F major chords, as shown in Table 2. Chord labels are short-hand for pitch collections and treating them as independent labels in evaluation metrics makes it much harder to decode where an algorithm is succeeding and failing since all errors are weighted equally.

The following accuracy metric for chord estimation evaluation can be added as a weighting to existing evaluation approaches and is applicable both in the audio and symbolic domains. [1] Let C be the number of predicted notes $\hat{y}$ in the ground truth correctly identified y

$$C = |y \cap \hat{y}| \tag{1}$$

Let I be the number of insertions (extra predicted notes) in the estimated chord that are not present in the ground truth.

$$I = |\hat{y} \backslash y| \tag{2}$$

Let A be the accuracy measurement for each chord estimate, calculated from C and I scaled between 0 and 1.

$$A = \frac{C - I + |y|}{2|y|} \tag{3}$$

Thus, A provides a combined measurement of which notes are correctly predicted and whether any additional notes (insertions), in excess of the number of notes in the ground truth chord, were predicted.

Using the example from above of F chord misestimated as either a d- chord or a G chord, we can see how this accuracy measurement captures the differences in the pitch content between the two estimates (using the pitch class content in Table 2). To calculate the accuracy for d-, A_d, we will compare the pitch classes of d-, {2,5,9}, to those of F, {0,5,9}.

$$C_d = |\{0, 5, 9\} \cap \{2, 5, 9\}| \tag{4}$$

$$C_d = 2 \tag{5}$$

$$I_d = |\{2, 5, 9\} \backslash \{0, 5, 9\}| \tag{6}$$

$$I_d = 1 \tag{7}$$

$$A_d = \frac{2 - 1 + |3|}{2|3|} \tag{8}$$

$$A_d = 0.66 \tag{9}$$

To calculate the accuracy for G, A_G, we will compare the pitch classes of G, {2,7,11}, to those of F, {0,5,9}.

$$C_G = |\{0, 5, 9\} \cap \{2, 7, 11\}| \tag{10}$$

$$C_G = 0 \tag{11}$$

$$I_g = |\{0, 5, 9\} \backslash \{2, 7, 11\}| \tag{12}$$

$$I_g = 3 \tag{13}$$

$$A_g = \frac{0 - 3 + |3|}{2|3|} \tag{14}$$

$$A_g = 0 \tag{15}$$

4. CONCLUSIONS

This proposed accuracy metric works directly on the pitch class information and can help to facilitate examinations of where chord estimation algorithms provide partially correct answers. Such an examination can lead to a more nuanced understanding of the algorithms and more efficient algorithm refinement. It can either be used as a weighting with existing evaluation approaches or further developed to replace the complex chord vocabularies currently in use.

[1] An implementation of the metric is available at `https://github.com/jcdevaney/chordEstimationMetric`.

5. REFERENCES

[1] J. Pauwels, K. O'Hanlon, E. Gómez, and M. Sandler, "20 years of automatic chord recognition from audio," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 54–63.

[2] B. McFee and J. P. Bello, "Structured training for large-vocabulary chord recognition," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 188–194.

[3] H. Koops, W. de Haas, J. Bransen, and A. Volk, "Chord label personalization through deep learning of integrated harmonic interval-based representations," in *Proceedings of the First International Conference on Deep Learning and Music, Anchorage*, 2017, pp. 19–25.

[4] E. J. Humphrey and J. P. Bello, "Four timely insights on automatic chord estimation." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 673–679.

[5] N. Condit-Schultz, Y. Ju, and I. Fujinaga, "A flexible approach to automated harmonic analysis: Multiple annotations of chorales by bach and prætorius." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 66–73.

[6] H. V. Koops, W. B. De Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019.

[7] H. Riemann, *Harmony simplified: Or the theory of the tonal functions of chords*. Augener & Company, 1895.

[8] W. Piston, *Harmony*. WW Norton New York, 1948.

[9] C. L. Krumhansl, *Cognitive foundations of musical pitch*. Oxford University Press, 1990.

[10] F. Lerdahl *et al.*, *Tonal pitch space*. Oxford University Press, USA, 2001.

[11] T. Crawford, J. Pickens, and G. Wiggins, "Dimensionality reduction in harmonic modeling for music information retrieval," in *International Symposium on Computer Music Modeling and Retrieval*. Springer, 2005, pp. 233–248.

[12] W. B. De Haas, R. C. Veltkamp, and F. Wiering, "Tonal pitch step distance: a similarity measure for chord progressions." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2008, pp. 51–56.

[13] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, vol. 5. Citeseer, 2005, pp. 304–311.

[14] J. Pauwels and G. Peeters, "Evaluating automatically estimated chord sequences," in *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 749–753.

[15] J. Devaney and C. Arthur, "Developing a structurally significant representation of musical audio through domain knowledge." in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[16] M. A. Kaliakatsos-Papakostas, A. I. Zacharakis, C. Tsougras, and E. Cambouropoulos, "Evaluating the general chord type representation in tonal music and organising GCT chord labels in functional chord categories." in *Proceedings of International Society for Music Information Retrieval (ISMIR) conference*, 2015, pp. 427–33.

[17] T. Carsault, J. Nika, and P. Esling, "Using musical relationships between chord labels in automatic chord extraction tasks," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2018.