THEME AND VARIATION ENCODINGS WITH ROMAN NUMERALS (TAVERN): A NEW DATA SET FOR SYMBOLIC MUSIC ANALYSIS

Johanna Devaney, Claire Arthur, Nathaniel Condit-Schultz, and Kirsten Nisula
School of Music, Ohio State University, USA
{devaney.12, arthur.193, condit-schultz.1, nisula.1}@osu.edu

ABSTRACT

The Theme And Variation Encodings with Roman Numerals (TAVERN) dataset consists of 27 complete sets of theme and variations for piano composed between 1765 and 1810 by Mozart and Beethoven. In these theme and variation sets, comparable harmonic structures are realized in different ways. This facilitates an evaluation of the effectiveness of automatic analysis algorithms in generalizing across different musical textures. The pieces are encoded in standard **kern format, with analyses jointly encoded using an extension to **kern. The harmonic content of the music was analyzed with both Roman numerals and function labels in duplicate by two different expert analyzers. The pieces are divided into musical phrases, allowing for multiple-levels of automatic analysis, including chord labeling and phrase parsing. This paper describes the content of the dataset in detail, including the types of chords represented, and discusses the ways in which the analyzers sometimes disagreed on the lower-level harmonic content (the Roman numerals) while converging at similar high-level structures (the function of the chords within the phrase).

1. INTRODUCTION

There are a wealth of musical scores in digitized form currently available. While the vast majority exist as images, a combination of hand encoding of the visual data and advances in optical music recognition (OMR) technology have increased the amount of symbolic music data available. Unfortunately, most of this data is unlabeled, limiting its utility in developing predictive systems for analyzing symbolically represented music. Accurately segmenting and labeling symbolic music data requires a higher level of musical expertise than can be reasonably obtained through crowd-sourcing platforms, like Mechanical Turk [1]. Even with expert-annotators, there is the challenge of ensuring that they all conform to the same conventions in labeling the data. In this regard, conforming to the analytic approach in a published textbook provides a measure of consistency for analyzing classical music.

This paper presents the Theme And Variation Encodings with Roman Numerals (TAVERN) datase [2], a new dataset of segmented and analyzed symbolic classical music. TAVERN consists of 27 theme and variations sets by Mozart and Beethoven, segmented into phrases and analyzed in terms of both Roman numeral chord labels and chord function. All of the pieces were analyzed in duplicate by different PhD-level music theory students and both the notes and analyses were encoded in Humdrum-related formats [9]. The dataset focuses on pieces in theme and variation form where the underlying harmony remains relatively constant across variations, while rhythmic and textural aspects of the music change. The utility of theme and variations in symbolic music analysis has been demonstrated in the case of folk songs [27, 28] and for both harmony [8, 14] and melody [5] in classical themes and variations. This is the first such dataset, however, that includes harmonic and functional data, facilitating the development of algorithms of automatic symbolic chord recognition and symbolic similarity, through a deeper understanding of the impact of texture on both of these tasks. This paper begins with a survey of existing symbolic music datasets, both annotated and unannotated, before describing in detail the annotation process and the contents of the dataset.

2. EXISTING DATASETS

As noted above, there is a growing number of unannotated symbolic music datasets available, many items of which are available in several collections. The most popular in MIR research are those that are hand-encoded and, to a certain degree, curated. This includes the KernScores dataset [22], which has more than 100,000 files in **kern format [9] from a range of styles from folk [23] to classical. A number of the kern score pieces are available in other datasets, such as the music21 corpus [3], which contains files in MusicXML [6] and **kern format. The music21 corpus also includes the Yale Classical Archives Corpus [29], which contains almost 9000 pieces/movements divided into vertical slices. The Yale corpus is also part of the ELVIS database [1] along with the Josquin Research

[1] http://www.mturk.com/

[2] http://getTAVERN.org

Project [3] and a number of smaller corpora of other Renaissance composers. While some datasets are focused on making printed versions of the musical scores available, they often supply symbolic data. For example, the Mutopia Project [4] contains not only PDFs of the scores but also hand-encoded Lilypond [5] and MIDI files. The Peachnote dataset [26] provides similar access to the Petrucci Music Library [6] by running OMR on the scanned scores, which typically has a higher error rate than hand encoding. Researchers have also made use of publicly available Band in a Box lead sheets, e.g, [4], and MIDI files, e.g., [15].

There is a much smaller number of harmonically annotated datasets. Temperley encoded the analyses from the *Tonal Harmony* textbook by Kostka and Payne [11] for his work on key finding [24] and examined statistical properties of harmony [25]. These encodings have been used by other researchers for evaluating symbolic chord recognition systems [12, 18]. The note data and annotations are available both in a format Temperley defined as "note files" [7] and as MIDI files (with the chord annotation inserted as lyrics). [8] The KSN harmonic annotations [10] provide Roman numeral labels with duration and inversion information for the Real World Computing (RWC) dataset [7] and have been used for modeling pitch structures in polyphonic music [19].

3. ANALYTIC APPROACH

TAVERN comprises 27 sets of theme and variations, 10 by Mozart and 17 by Beethoven (listed in Table 1). The Beethoven set is nearly complete, with 18 of his 20 theme and variation sets included (Opus 35 was excluded because of the inclusion of a fugue in the piece and Wo0 79 was excluded because it included only 5 variations, which was below our 6 variation minimum). The Mozart set is less complete: due to time and resource restrictions, we temporally sampled variations across his career (leaving out K. 24, 54, 180, 264, 352, 460, 500). Going forward we plan to analyze and include these variations in the dataset once additional resources become available.

The pieces have been analyzed in duplicate by multiple expert-annotators using the hierarchical model of harmony defined in [13] that includes both Roman numeral and function labels, specifically a variant of functional analysis known as the 'Phrase Model'. Section 3.1 provides some background on the 'Phrase Model' in general and Section 3.2 describes the annotation process.

3.1 Phrase Model

Phrases are complete musical statements built from an ordered presentation of three harmonic functions and ending with a cadence. One way of analyzing phrases is in terms of functions. The tonic function at the beginning of a phrase serves to establish the tonal centre, and at the end of a phrase to signal its return. The pre-dominant function prepares for the arrival of the dominant function, which sets up an opposition to tonic. The tension created by the movement to the dominant is ultimately resolved by a return to tonic. A phrase typically contains all three harmonic functions, but may contain just tonic and dominant. The cadences may close with the dominant function (termed a half cadence) or return to the tonic function (termed an authentic or deceptive cadence, depending on the chords used). Ideas about functional harmony can be found in Rameau [20], although the specification of the terms tonic, pre-dominant, and dominant were not defined until the late nineteenth century by Riemann' [21]. We have included function labels in addition to the Roman numeral labels because we believe that they are essential in developing and testing hierarchical models of harmony. Since function harmony has some limitations for music outside of the Classical era, we focused this dataset on Mozart and early-mid career Beethoven pieces.

The 'Phrase Model' is a contemporary adaption of Riemann's thinking and is defined in several textbooks. For the purposes of this project, we followed the specifics laid out in *The Complete Musician* by Steven Laitz [13]. Gen-

Composer	Piece	# Variations
Mozart	K.25	7
	K.179	12
	K.265	12
	K.353	12
	K.354	12
	K.398	6
	K.455	10
	K.501	12
	K.573	9
	K.613	8
Beethoven	WoO 63	9
	WoO 64	6
	WoO 65	24
	WoO 66	13
	WoO 68	12
	WoO 69	9
	WoO 70	6
	WoO 71	12
	WoO 72	8
	WoO 73	10
	WoO 75	7
	WoO 76	8
	WoO 77	6
	WoO 78	7
	WoO 80	32
	Opus 34	6
	Opus 76	6

Table 1. Summary of the sets of themes and variations in the data set.

[3] http://jrp.ccarh.org/
[4] http://www.mutopiaproject.org
[5] http://www.lilypond.org
[6] http://imslp.org
[7] http://theory.esm.rochester.edu/temperley/kp-stats/index.html
[8] http://www.cs.northwestern.edu/~pardo/kpcorpus.htm

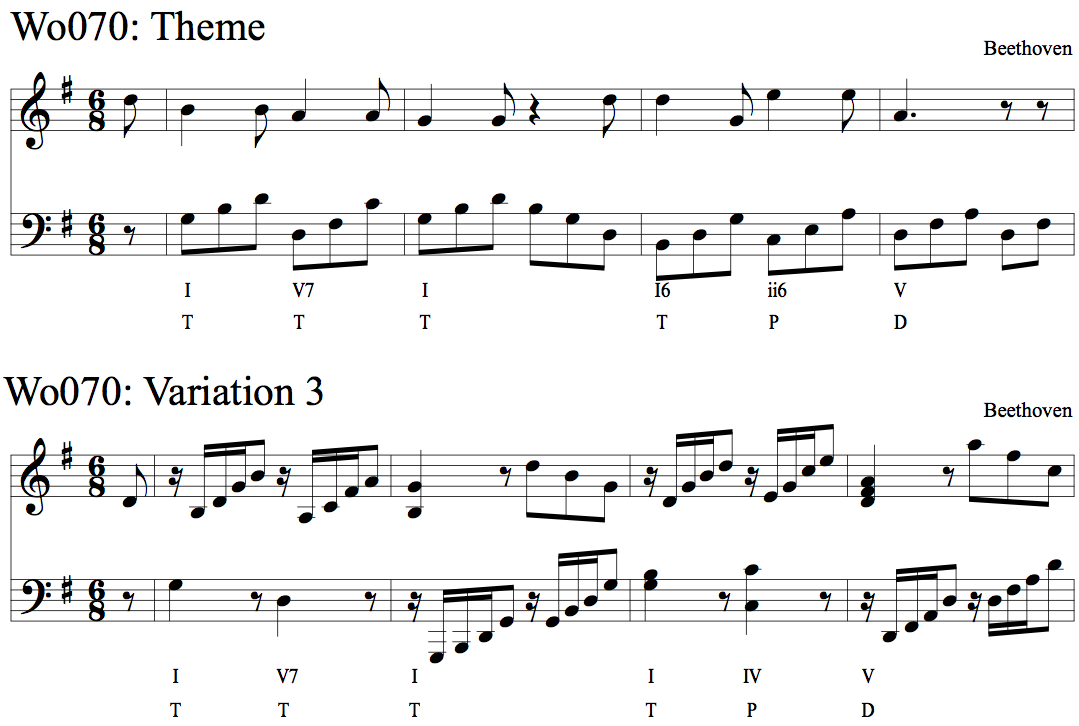

Figure 1. Example of a theme and variation from the dataset with harmonic analyses marked, note the similarity in the harmonic structure and the differences in the texture.

erally, the majority of I and iii chords (i and III in the minor mode) have a tonic function, although inversions of these chords may have other function, such as I_4^6 functioning as dominant, depending on their harmonic context. vi (or VI) chords may have either a tonic or pre-dominant function, while ii or IV (ii^o or iv) chords are typically pre-dominant. V and vii^o chords are typically assigned a dominant function, except for when their inversions occur in passing or neighbor contexts with I or vi chords in a tonic function. An example of the 'Phrase Model' analytical approach is shown in Figure 1. In the Theme, the Roman numerals I-V^7-I-I^6 are assigned a tonic function, with the V^7 in the first bar functioning as a ornamentation of the surrounding I chords, rather than having a dominant function. The ii^6 chord has a pre-dominant function and the V chord has a dominant function. Since the phrase ends on the dominant function, rather than returning to the tonic function, it ends with a half cadence. The variation has a similar structure, with the first 2.5 measures having a tonic function, the second half of the third measure having a pre-dominant function (albeit with a IV chord instead of ii^6 chord), and the fourth measure having a dominant function.

3.2 Annotators

The annotators are three PhD-level music theory students, who each have spent at least two years teaching the harmonic analysis technique described in Section 3.1 to undergraduate students within the same curricular framework. Thus the annotators are intimately familiar with the workings of Laitz's version of the 'Phrase Model' and its analytic conventions, ensuring a common interpretation across the annotators on these conventions. Each of the theme and variations sets was analyzed by two annotators, with the annotators analyzing 18 theme and variations sets each. The annotators worked independently, dividing each of the themes and variations into phrases on their own and analyzing each phrase both in terms of Roman numerals and phrase-level function. In cases where there was disagreement between the annotators, a third annotator reviewed the analyses and sided with one interpretation. The adjudicated version of the analysis was then joined with the note data, as described in Section 4.1. On occasion, the analyzers would disagree on the Roman numerals while still agreeing about the function of the chords, an example of which is discussed in Section 4.2 We believe that points of disagreement between the trained annotators are an interesting source of information, particularly if chord recognition algorithms run into accuracy issues in the same situations, and so we are also releasing the individual annotations in addition to the adjudicated data.

4. DATASET DETAILS

4.1 Encoding Format

The musical scores of pieces were converted from publicly available MIDI files sourced online. The MIDI files were less error-prone than running OMR on printed scores of the pieces, but still required some manual correction. In the correction process, the MIDI files were first converted into **kern format after which the errors were hand-

```
!!!COM: Beethoven
!!!OTL: 6 Variationen uber "Nel cor
!!!piu non mi sento" von G. Paisiello
!!!Variation: Theme a
**func  **harm  **kern  **kern
*M6/8   *M6/8   *M6/8   *M6/8
*G:     *G:     *G:     *G:
*tb8    *       *       *
8D      8V      8r      8dd
=       =       =       =
4.T     4.I     8GL     4b
.       .       8B      .
.       .       8dJ     8b
4.T     4.V7    8DL     4a
.       .       8F#     .
.       .       8cJ     8a
=       =       =2      =2
2.T     2.I     8GL     4g
.       .       8B      .
.       .       8dJ     8g
.       .       8BL     4r
.       .       8G      .
.       .       8DJ     8dd
=       =       =3      =3
4.T     4.I6    8BBL    4dd
.       .       8D      .
.       .       8GJ     8g
4.P     4.ii6   8CL     4ee
.       .       8E      .
.       .       8AJ     8ee
=       =       =4      =4
4.D     4.V/V   8DL     4.a
.       .       8F#     .
.       .       8AJ     .
4D      4V      8DL     8r
.       .       8F#     8r
*-      *-      *-      *-
```

Figure 2. Example of the encoding format for the theme in Figure 1. The leftmost column contains the function labels, the second one contains the harmonic labels, and the remaining columns contain the notes. Dots indicate that a label is continued from a previous row while elements of another spine change.

corrected in reference to public domain scores available in the Petrucci Music Library (namely 19th century publications from Breitkopf & Hartel [2,16,17]). In the correction process, ornamentation and grace notes were removed in order to simplify the data. In addition to pitch and duration information, **kern format allows for information about slurs and stem directions to be encoded. Where this information was encoded in the MIDI data, it was converted into the **kern data.

The analyses were encoded as separate spines and then joined with the **kern data. For the Roman numeral analysis the existing **harm representation [9] was used. In this format, the labels are the same as standard Roman numeral labels except that the inversions are marked with the letters a (for first inversion, typically notated as ^{6} for triads or $\frac{6}{5}$ for seventh chords), b (for second inversion, $\frac{6}{4}$ or $\frac{4}{3}$), and c (for third inversion, $\frac{4}{2}$) in order to maintain consistency for the number of character used to indicate inversions. We developed a new format, named **func, for the function encoding, which simply consists of the labels T (tonic), P (pre-dominant), and D (dominant). Thus each file consists of one **func spine, one **harm spine and a number of **kern spines, each of which corresponds to one staff in the piano score. An example of a file, corresponding to the upper scores in Figure 1, is shown in Figure 2. Each file in the dataset represents one phrase, with measure numbers marked in reference to the entire piece. This allows for the phrases across the corresponding theme and variations to be easily recombined into a single piece while at the same time providing an indication of where each phrase begins and ends. The files are readable by Humdrum, a MATLAB parser for the files is currently available on github [10] and extensions to the music21 Humdrum parser will be available shortly. We have also generated audio versions of each file from the symbolic data via MIDI.

[9] http://www.humdrum.org/Humdrum/representations/harm.rep.html

4.2 Theme and Variations Form

All 27 of theme and variations sets in TAVERN are in 'sectional' form, meaning that all of the themes and variations are tonally-closed distinct units. Within the sets, the harmony remains relatively constant across the theme and variations, while the theme's melody is embellished in the variations. Additional musical interest is created through changes in rhythm, tempo, texture, key, and mode. There are some inconsistencies in the harmonies across related themes and variations, but these are typically substitutions of different chords with the same harmonic function. An example of this is present in Figure 1, where the ii^6 chord in the penultimate measure of the theme is substituted with a IV chord in the variation. However, ii^6 and IV share two common notes (the 4th and 6th scale degrees) and a common function (P), meaning that this substitution has very little harmonic impact.

In total, the dataset consists of 1060 phrases. Of these, 66 phrases occur as codas to isolated variations, so for these phrases there is no corresponding phrase in the related theme or variations. These have been included for the purposes of completeness. Of the 1060 phrases, 917 of the phrases are in the major mode, with the remaining 143 being in the minor mode. Seven different major and minor keys are occur in the dataset: A, B flat, C, D, E flat, F, G). Within the phrases there are 290 unique sonorities (counting each inversion as a separate sonority), this includes both diatonic chords and applied chords. A tally of the top 40 unique chords with the highest number of occurrences (at least 25) is shown in Figure 3, along with the number of times that each chord occurs in each function. In addition to highlighting the large number of chords that are annotated in the dataset, Figure 3 also demonstrates the utility of annotating function labels by showing that most of the chord inversions have two if not three possible functions (depending on the context in which they occur). This highlights the need for such labelled data in order to learn these contexts, rather than simply relying on rule-based systems.

The relatively large proportion of non-standard tonic chords with a tonic function in Figure 3 (e.g., ii, IV, V, vii^o) are a result of "embedded phrases" within the tonic function in some of phrases [13]. An example of this is shown in the **comments spine of Annotator Two's analysis of

[10] https://github.com/jcdevaney/TAVERN

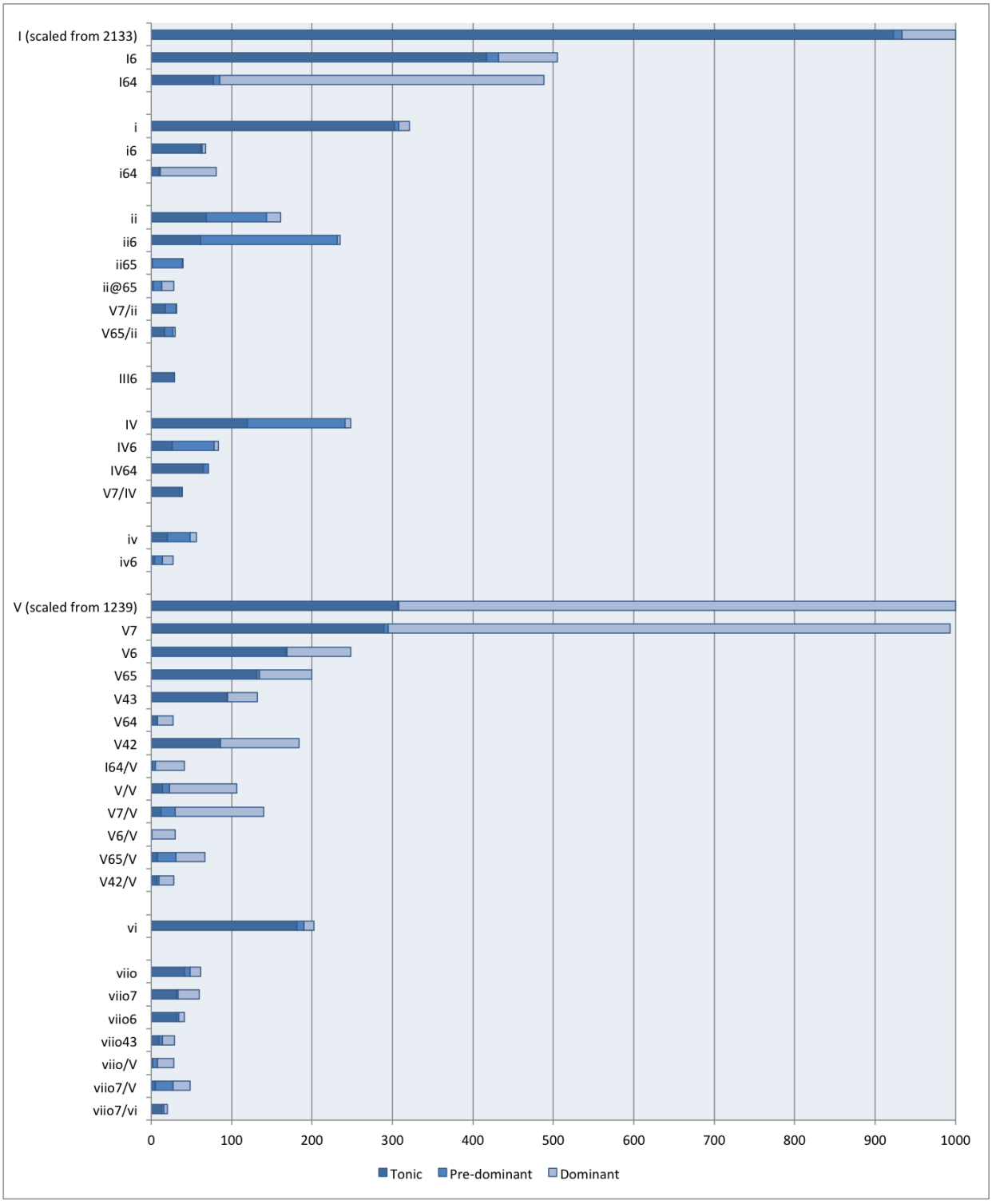

Figure 3. A tally of the number of times each of the top forty unique chords occurs in the dataset in regards to the function (Tonic, Pre-dominant, Dominant) in which they occur. The data for the I and V chords are shown the number of occurrences per 1000, scaled from their total number of occurrences (2133 and1239 occurrences, respectively). This was done to facilitate the readability of the figure. The chords are grouped, from top to bottom, by the scale degree of their root note (or in the case of applied chords, the diatonic scale degree which functions as their relative tonic). Within each chord group, the chords are ordered by inversion followed by occurrences of applied dominant chords on that scale degree.

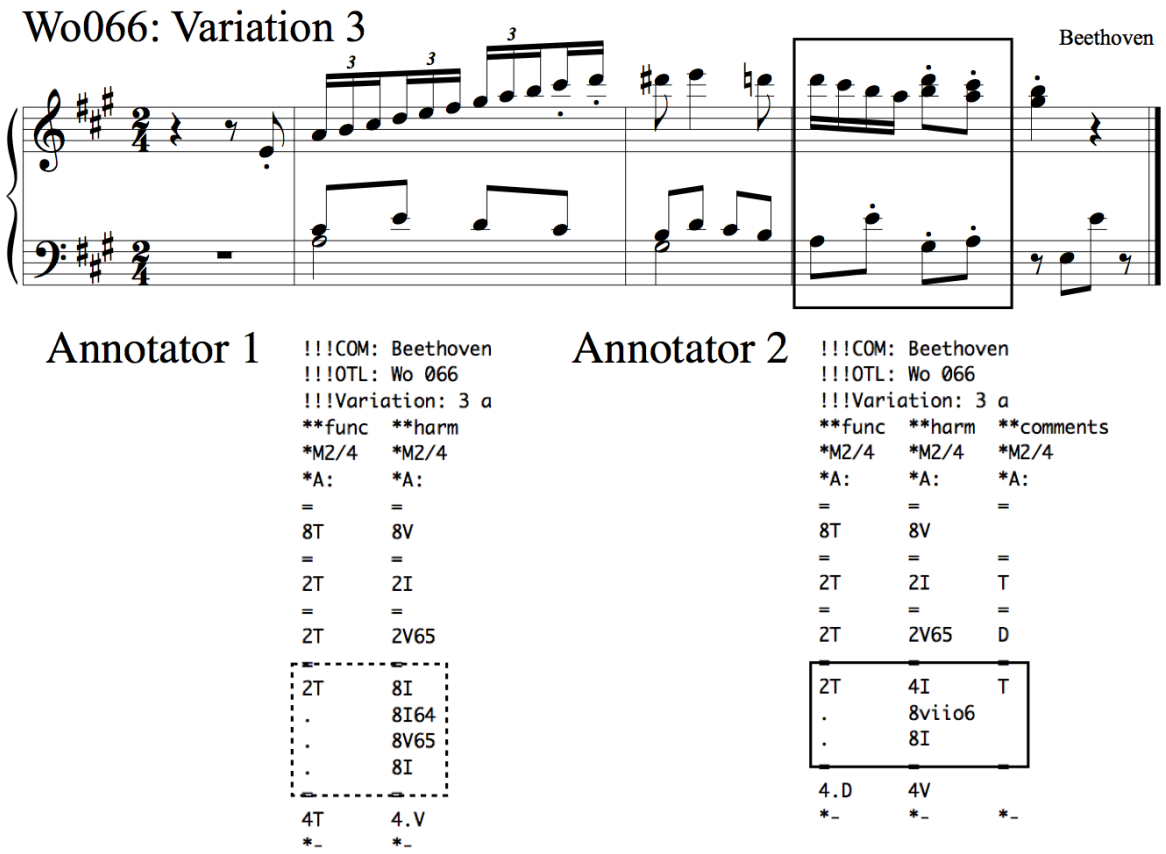

Figure 4. An example of a phrase where the two annotators disagreed on specific chord labels. In the third measure (marked with a box), Annotator 1 analyzed the measure as 'I- I^6_4- V^6_5-I' while Annotator 2 analyzed the measure as 'I-vii^{o6}-I'. The adjudicating annotator sided with Annotator 2 because in this context ' vii^{o6}' label is a complete chord. ' V^6_5', despite being technically correct, is less desirable because the root of the chord (E) is missing. Annotator 2's analysis also demonstrates the nomenclature of 'embedded phrases', which are marked when there is a low-level 'T-P-D-T' or 'T-D-T' pattern within the main T function that does not result in a cadence. Where applicable, 'embedded phrase' analyses are available in the individual annotators' files in the **comments spine.

musical phrases reproduced in Figure 4. Instances of embedded phrases are not included in the main database files, but are available in the individual annotator's files that are also released as part of TAVERN. Figure 4 also provides an example where the two annotators agreed on the overall harmonic function, but disagreed on the specific Roman numerals (as seen in the different analyses for measure 3). Ultimately, in this case, a third annotator determined the second annotator's analysis to be superior both because the chord labels described complete chords and because it better mirrored the harmonic activity in the corresponding phrases in the related theme and variations.

5. CONCLUSIONS

This paper has presented TAVERN, a new dataset of 27 harmonically annotated theme and variations piano pieces by Mozart and Beethoven that will facilitate research on symbolic chord recognition and similarity in symbolic music. Each musical phrase in the dataset is encoded as a separate file. The note information is encoded in **kern format, the Roman numerals in **harm format, and the harmonic function of each Roman numeral label in the newly defined **func format.

This dataset will be useful for systematically evaluating the effect of textural changes on symbolic chord recognition algorithms since the consistency of harmonic materials and melodic frame across each theme and variations set occurs against a wide range of musical textures. Also, the segmentation of the pieces into phrases can facilitate the development and evaluation of algorithms for musical structure analysis. In addition to the symbolic music data, MIDI-generated audio files are available. In the future, we plan to use score-audio alignment to generate a mapping between the symbolic data and public-domain recordings of real piano performances, extending the utility of this dataset to include audio chord recognition research.

6. ACKNOWLEDGMENTS

This work was supported by the Google Faculty Research Award program.

7. REFERENCES

[1] Christopher Antila and Julie Cumming. The vis framework: Analyzing counterpoint in large datasets. In *Proceedings of ISMIR*, pages 71–6, 2014.

[2] Ludwig van Beethoven. *Variationen für das Pianoforte*, volume Serie 17 of *Ludwig van Beethovens Werke*. Breitkopf & Härtel, Leipzig, DE, 1862-90.

[3] Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of ISMIR*, pages 637–42, 2010.

[4] W. Bas De Haas, Martin Rohrmeier, Remco C Veltkamp, and Frans Wiering. Modeling harmonic similarity using a generative grammar of tonal harmony. In *Proceedings of the ISMIR*, pages 549–54, 2009.

[5] Mathieu Giraud, Ken Déguernel, and Emilios Cambouropoulos. Fragmentations with pitch, rhythm and parallelism constraints for variation matching. In *Sound, Music, and Motion*, pages 298–312. Springer, 2014.

[6] Michael Good. MusicXML for notation and analysis. *The virtual score: representation, retrieval, restoration*, 12:113–24, 2001.

[7] Masataka Goto, Hiroki Hashiguchi, Takuya Nishimoto, and Ryuichi Oka. Rwc music database: Popular, classical, and jazz music databases. In *Proceedings of ISMIR*, pages 287–288, 2002.

[8] Keiji Hirata, Satoshi Tojo, and Masatoshi Hamanaka. Cognitive similarity grounded by tree distance from the analysis of k. 265/300e. In *Sound, Music, and Motion*, pages 589–605. Springer, 2014.

[9] David Huron. *The Humdrum Toolkit: Reference Manual*. CCARH, Menlo Park, California, 1995.

[10] Hitomi Kaneko, Daisuke Kawakami, and Shigeki Sagayama. Functional harmony annotation database for statistical music analysis. In *Proceedings of the ISMIR (Late Breaking Demo)*, 2010.

[11] S. Kostka and D. Payne. *Tonal Harmony: With an Introduction to Twentieth Century Music*. McGraw-Hill, New York, NY, 2008.

[12] Pedro Kröger, Alexandre Passos, Marcos Sampaio, and Givaldo De Cidra. Rameau: A system for automatic harmonic analysis. In *Proceedings of the International Computer Music Conference*, pages 273–281, 2008.

[13] Steven G. Laitz. *The Complete Musician*. Oxford University Press, Oxford, 3rd edition edition, 2011.

[14] Alan Marsden. Recognition of variations using automatic schenkerian reduction. In *Proceedings of ISMIR*, pages 501–506, 2010.

[15] Matthias Mauch and Simon Dixon. A corpus-based study of rhythm patterns. In *Proceedings of ISMIR*, pages 163–168, 2012.

[16] Wolfgang Amadeus Mozart. *Für ein und zwei Pianoforte zu vier Händen*. Wolfgang Amadeus Mozarts Werke, Serie XIX. Breitkopf & Härtel, Leipzig, DE, 1878.

[17] Wolfgang Amadeus Mozart. *Variationen für das Pianoforte*. Wolfgang Amadeus Mozarts Werke, Serie XXI. Breitkopf & Härtel, Leipzig, DE, 1878.

[18] B. Pardo and W. Birmingham. Algorithms for chordal analysis. *Computer Music Journal*, 26(2):27–49, 2002.

[19] Stanislaw Andrzej Raczynski, Emmanuel Vincent, and Shigeki Sagayama. Dynamic bayesian networks for symbolic polyphonic pitch modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1830–1840, 2013.

[20] Jean-Phillipe Rameau. *Treatise on Harmony*. Dover, Toronto, ON, 1722.

[21] Hugo Riemann. *Harmony Simplified; or, the Theory of the Tonal Functions of Chords.* Augener, London, 1896.

[22] Craig Stuart Sapp. Online database of scores in the humdrum file format. In *Proceedings of ISMIR*, pages 664–665, 2005.

[23] Helmut Schaffrath and David Huron. *The Essen folksong collection in the humdrum kern format*. CCARH, Menlo Park, CA, 1995.

[24] David Temperley. *A Bayesian Approach to Key-Finding*, volume 2445 of *Lecture Notes in Computer Science*, pages 195–206. Springer Berlin Heidelberg, 2002.

[25] David Temperley. A unified probabilistic model for polyphonic music analysis. *Journal of New Music Research*, 38(1):3–18, 2009.

[26] Vladimir Viro. Peachnote: Music score search and analysis platform. In *Proceedings of ISMIR*, pages 359–362, 2011.

[27] Anja Volk, WB Haas, and P Kranenburg. Towards modelling variation in music as foundation for similarity. In *Proceedings of the International Conference on Music Perception and Cognition*, pages 1085–1094, 2012.

[28] Anja Volk and Peter van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 16(3):317–339, 2012.

[29] Christopher White and Ian Quinn. The Yale-Classical Archives Corpus. In *Proceedings of the International Conference on Music Perception and Cognition*, page 320, 2014.