DESIGN AND CREATION OF A LARGE-SCALE DATABASE OF STRUCTURAL ANNOTATIONS

Jordan B. L. Smith[1], J. Ashley Burgoyne[2], Ichiro Fujinaga[2], David De Roure[3], and J. Stephen Downie[4]
[1]University of Southern California, [2]McGill University,
[3]University of Oxford, [4]University of Illinois at Urbana-Champaign
jordans@usc.edu, ashley@music.mcgill.ca, ich@music.mcgill.ca,
david.deroure@oerc.ox.ac.uk, jdownie@illinois.edu

ABSTRACT

This paper describes the design and creation of an unprecedentedly large database of over 2400 structural annotations of nearly 1400 musical recordings. The database is intended to be a test set for algorithms that will be used to analyze a much larger corpus of hundreds of thousands of recordings, as part of the Structural Analysis of Large Amounts of Musical Information (SALAMI) project. This paper describes the design goals of the database and the practical issues that were encountered during its creation. In particular, we discuss the selection of the recordings, the development of an annotation format and procedure that adapts work by Peeters and Deruty [10], and the management and execution of the project. We also summarize some of the properties of the resulting corpus of annotations, including average inter-annotator agreement.

1. INTRODUCTION

The Structural Analysis of Large Amounts of Musical Information (SALAMI) project is a musicological endeavour whose goal is to produce structural analyses for a very large amount of music—over 300,000 recordings. Here structure refers to the partitioning of a piece of music into sections and the grouping together of similar or repeated sections. These sections usually correspond to functionally independent sections, such as the "verse" and "chorus" sections of a pop song, the "exposition" and "development" of a sonata—or, at a shorter timescale, the exposition's "main theme," "transition," and "secondary theme" groups.

The recordings in the SALAMI corpus represent an enormous range of genres, from klezmer to top-40 pop, and a variety of sources, including professional studio recordings and audience-recorded live sessions. The SALAMI dataset, which will be made freely available, could be of great service to music theorists, musicologists, and other music researchers, since determining the form of an individual piece of music is generally a time-consuming task. The SALAMI dataset could facilitate large-scale studies of form, which presently are relatively uncommon.

Because of the value of knowing the structure of pieces of music, the pursuit of algorithms that produce structural descriptions automatically is an active area of research. (For a review see [9].) The SALAMI project plans to use a selection of these algorithms to analyze its hundreds of thousands of recordings. However, before these algorithms can be used, it is necessary to validate their performance on the vast array of genres represented. This demands the creation of a human-annotated ground truth dataset. The design and creation of a large database such as the SALAMI test set raises many methodological issues relating to the choice of music, annotation format, and procedure. This paper explains the issues involved and the decisions we made to address them.

The next section of this work summarizes the content and contributions of several existing corpora of structural annotations, as well as important recent research on the annotation process itself [1, 10]. Section 3 describes the creation of the SALAMI test set, including the corpus selection, the annotation format used, and the recommended workflow. Some properties of the resulting dataset are presented and discussed in Section 4.

2. PRIOR WORK

2.1 Existing collections

SALAMI requires a database that includes a significant amount of popular, jazz, classical, and world music.[1] However, most previous collections of annotations only consider popular music. Three of the largest existing databases of annotations are *TUTstructure07* [13] (557 annotations), compiled at Tempere University of Technology (TUT) and containing mainly popular music; annotations for the Beat-

[1] These four genre labels should be understood in their broadest sense, so that together they encompass all music. Thus "classical" refers to all Western art music; "popular" refers to most modern commercial music, including The Cure and Autechre; and so forth for "jazz" and "world."

les studio catalogue created by Alan Pollack and synchronized independently by two groups, including the Centre for Digital Music [5] (180 annotations); and the AIST Annotation set [4] that accompanies the RWC Music Database (285 annotations). The RWC set is approximately half popular music, and one quarter each jazz and classical, with an additional few world music pieces, but for many of the jazz and classical pieces only the "chorus" sections are indicated.

2.2 Annotation formats

Nearly all previous corpora of annotations have used the same straightforward annotation format. Pieces are segmented into non-overlapping sections, and every section is given a single label, such as "intro" or "chorus," to indicate which are similar to or repetitions of one another. The labels also suggest the musical role or function of each section. In some corpora, such as the Beatles annotations [5], labels may indicate instrumentation (e.g., "verse_guitar") or variations on a section (e.g., "verse_with_ending").

2.2.1 Issues with previous formats

As pointed out in Peeters and Deruty [10], this conflation of musical similarity, function, and instrumentation is problematic. For instance, a song's "outro" may use the same music as an earlier "transition," but labelling them as such fails to record their similarity. Contrariwise, a section with a single function may be musically heterogenous, as with an extended two-part introduction. Peeters and Deruty also criticized the large, seemingly unconstrained vocabularies used in certain collections of annotations. Consider again the Isophonics Beatles annotations [5]: of the 146 unique labels, 95 are used just once. Single-use labels may be informative to a human inspecting the annotation, where their meaning is understandable in context (e.g., "intro_redux," "verse_(slow)"), but having too many unique labels is less useful when the annotations are being used by a machine. Another drawback of the standard annotation format is that it only describes the structure at a single timescale. One of the most important attributes of musical structure is that it is perceived hierarchically, and it would be ideal to capture some of this information in an annotation.

2.2.2 An alternative format

Peeters and Deruty proposed an alternative annotation format intended to resolve these problems. The format uses a restricted vocabulary of 19 labels, each of which addresses one of three aspects of a piece's structure: either musical similarity, musical role, or instrument role. In their format, musical similarity is indicated by labelling every portion of a piece as one of five "Constitutive Solid Loops" (CSLoops). (If more than five are required, a sixth CSLoop is used, although the format does not imply that all sections labelled with this last label are similar.) Function labels are optional and are restricted to "intro/outro," "transition," "chorus," and "solo." Instrumentation labels indicate whether a primary or supporting melodic voice is present.

Peeters and Deruty's format also creatively incorporates some hierarchical information about the structure. Two markers, "V1" and "V2," divide CSLoops; the first indicates that the musical segments on either side of the marker are similar, the second that they are dissimilar.

2.3 Annotation procedures

Unlike pitch and, to a large extent, beat, the perception of structure is a highly subjective phenomenon, and it is common for two listeners to disagree on the form of a piece of music. It is therefore challenging to develop an annotation procedure that, while perhaps not being objective, maximizes the repeatability of the results. Note that since a structural analysis records a listener's creative interpretation as much as her perception, objectivity is arguably an impossible goal for annotations.

One approach is to treat the creation of annotations as a perceptual experiment, and simply have multiple subjects listen to a piece and press a button whenever they perceive a structural boundary. Such data were collected by [2], who noted that listeners generally agreed on the placement of boundaries that they judged most salient. These boundaries were used as a type of "ground truth" by the authors to evaluate the success of some computational models at estimating boundaries.

Bimbot et al. [1] managed to obtain a degree of repeatability by precisely specifying an annotation procedure. They defined the musical criteria and similarity judgements an annotator should use in order to estimate boundaries. (The task of labelling the segments remains future work.) They reported that with their procedure, annotations were very consistent across annotators and over time. An annotator's goal is to decompose a piece into "autonomous and comparable blocks." Autonomy means that whether a block stands alone or is looped continuously, the result should be musically acceptable. Two blocks may be comparable if they have the same duration in beats, are interchangeable, or are similar with respect to their temporal organization.

3. DESCRIPTION OF THE SALAMI TEST SET

We developed a new corpus of annotations using a unique annotation format to address the goals of the SALAMI project. To ensure that the corpus was useful as an evaluation test set for SALAMI, the main design consideration was for the corpus to cover as wide a variety of musical genres as possible. For the annotations to be musicologically useful, the design goals for the annotation format were to have musical similarity, function, and lead instrumentation described independently, and for the annotations to reflect the hierarchical nature of musical structure. Finally, the format and the procedure should allow annotations to be produced

quickly, to minimize cost, but be flexible enough to handle works from a wide range of genres, all while aiming for high inter-annotator agreement. With these design considerations in mind, we conducted a survey of previous corpora of annotations and existing annotation techniques. Based on this survey and on our own experimentation with different approaches, we settled on the corpus, format, and procedure outlined in this section.

3.1 Contents of SALAMI test set

The first step in designing the corpus was deciding what to put in it. One of SALAMI's priorities was to provide structural analyses for as wide a variety of music as possible, to match the diversity of music to be analyzed by the algorithms. In addition to popular music, the SALAMI test set should pay equal attention to classical, jazz, and non-Western music known colloquially as "world" music. To ensure a diversity of recording formats, we also emphasized the inclusion of live recordings. The final composition of the database is shown in Table 1.

A secondary goal of the SALAMI test set was to be able to compare our annotations with those of previous data sets. We thus duplicated some previous work: our test set presently includes 97 and 35 recordings from the RWC and Isophonics data sets, respectively. Note that these recordings are all single-keyed (i.e., annotated by a single person), whereas most of the SALAMI test-corpus is double-keyed (analyzed by two independent annotators). Double-keying provides useful information but is more expensive. Single-keying some entries seemed to be a reasonable compromise given that other groups had already annotated these pieces.

Class	Double keyed	Single keyed	Total	Percentage
Classical	159	66	225	16%
Jazz	225	12	237	17%
Popular	205	117	322	23%
World	186	31	217	16%
Live music	273	109	382	28%
Total	1048	335	1383	100%

Table 1. Number of pieces of each class in the SALAMI test set. Single and double keying refers to the number of annotators (2 or 1, respectively) who independently analyzed each song.

Selecting songs for the corpus by hand would be time-consuming and would introduce unknown methodological bias. However, selecting songs randomly from most sources would result in a corpus heavily skewed toward popular music. To resolve this, most of the recordings were collected from Codaich [7], a large database with carefully curated metadata, including over 50 subgenre labels. This enabled us to enforce good coverage of genres while still choosing individual pieces randomly. The remainder of the test set was collected randomly from the Live Music Archive [6]. Unfortunately, metadata for these recordings is inconsistent and a distribution by genre could not be enforced. The majority appears to be popular and jazz music.

3.2 Annotation format

We developed a new annotation format that takes after the format devised by Peeters and Deruty in many important ways: we borrow their tripartite distinction between labels that indicate musical similarity, function, and instrumentation, and like them we also strictly limit the vocabulary of function labels. However, we have made several modifications to suit SALAMI's unique needs and more musicological focus. The labels in each of the three layers are described in the following three sections. An example annotation is shown in Figure 1.

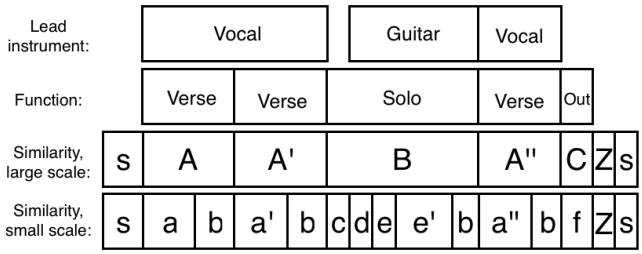

Figure 1. Example to illustrate proposed format.

3.2.1 Musical similarity track

The musical similarity track includes two layers at different timescales, each identifying which portions of the piece use similar musical ideas. The large-scale layer uses uppercase letters as labels ("A," "B," etc.) and the small-scale layer uses lowercase letters ("a," "b," etc.). The use of letter labels mimics the familiar music-theoretical approach. Every portion of a recording in both large- and small-scale layers must be assigned a letter label. The format specification allows any number of lowercase or uppercase letters to be used (the labels "aa," "ab," and so on may be used if the alphabet is exhausted). However, for the large-scale layer, annotators were instructed to prefer to use five or fewer distinct uppercase labels per recording. This preference rule does not express an assumption that there are five or fewer distinct musical ideas in any recording. Rather, it is intended to guide the annotator toward a certain level of abstraction. This direction proved useful when annotating works that are less clearly organized into distinct sections, such as through-composed pieces. It also helps when annotating works such as sonatas that may be organized into sections, but where these sections are not musically homogenous and may include several distinct musical ideas.

Two additional special labels indicate silence ("silence") and non-music, such as applause or banter in a live recording ("Z"). We also allow letter labels to be inflected by the prime symbol (′) to indicate a section that is evidently similar to another, but that is judged to be substantially varied. Similarity judgements are inherently subjective and imprecise, and the prime symbol is a useful way of acknowledging this. It allows the annotator to faithfully record his interpretation, while allowing future users to easily adapt the labels according to their needs. For instance, depending on the application, a user may excise the prime markers (so that "a" and "a′" are both relabelled as "a") or to treat variations as distinct sections (so that "a′" would be reassigned a letter label different from "a").

3.2.2 Function track

The second track in the annotation format contains the music function labels, which all must be drawn from our strict vocabulary of 20 labels. Unlike the letter labels, it is not mandatory that every portion of a piece receive a function label. The vocabulary is listed in Table 2, separated into various relevant categories. The instrumental, transition, and ending groups are all synonym groups. Note that in the ending group, the label "fadeout" is a special label that can occur in addition to any other label. For example, if the piece fades out over a repetition of the chorus, then the last section may be given both labels: "chorus" and "fadeout." Full definitions for each term are specified in our Annotator's Guide, available online [11].

Basic group	intro, verse, chorus, bridge
Instrumental	instrumental, solo
Transition	transition, pre-chorus, pre-verse, interlude
Genre-specific	head, main theme, (secondary) theme
Form-specific	exposition, development, recapitulation
Ending	outro, coda, fadeout
Special labels	silence, end

Table 2. List of permitted function words in proposed annotation format.

Note that some of the labels are genre-specific alternatives to others: for example, the "head" in a jazz song is analogous to a "chorus" in a pop song or, sometimes, a "main theme" in a classical piece. Also, together, the terms "exposition," "development," and "recapitulation" are specific to sonata form and may in special cases be used to annotate a third level of structural relationships at a timescale larger than the large-scale similarity labels. However, "development" also has wider applicability: it may be used to indicate the function of a contrasting middle section, which is relevant in many contexts, from various classical genres to progressive rock. Additionally, some subsets of the vocabulary can function as synonym-groups that can be collapsed into a single function label if desired. For example, while our Annotator's Guide defines a relatively subtle distinction between "pre-chorus," "pre-verse," "interlude," and "transition" sections, they are all synonyms of "transition." This approach allows annotators to err on the side of precision, while enabling future users of the data to ignore distinctions that are unneeded.

3.2.3 Lead instrument track

The final track in the annotation format indicates wherever a single instrument or voice takes on a leading, usually melodic role. The labels in this track are simply the names of the leading instruments, and hence the vocabulary is not constrained. Also, unlike the other tracks, lead instrument labels may potentially overlap, as in a duet. Note that as with the function track, there may be portions of the recording with no lead instrument label, if no instrument fulfills a leading role.

Note that in the written format devised for this project, the boundaries delineating the small-scale similarity segments are the only available boundaries when annotating the function and lead instrumentation tracks. Again, this helps orient annotators to an appropriate level of abstraction, and relieves them of too painstakingly indicating the instrumentation changes.

3.3 Annotation procedure

The annotators used the software Sonic Visualiser [3] to audition and annotate the pieces. Sonic Visualiser's keyboard commands allow one to insert and label boundaries quite quickly. We suggested the following workflow: first, listen through the song and mark a boundary whenever a structural boundary is perceived. Second, listen to the piece again, adjusting boundaries and adding lowercase labels. Third, add the uppercase and function labels, and finally add the lead instrument labels. While we found this workflow to be efficient and straightforward, we did not demand that annotators follow this or any other specific workflow.

3.4 Project realization

The annotation format and data collection took place over the course of 10 months. First, previous annotation formats and databases of annotations were researched. Potential annotation formats were devised and tested by the project leaders, and a tentative format was set at the end of two months. Next, candidate annotators were trained in the annotation format and in the Sonic Visualiser environment. Eight successful candidates were hired, all pursuing graduate studies in either Music Theory or Composition, and data collection began the following week. Because the annotation format had not been tested on a significant scale before work began in earnest, the first six weeks of data collection were conceived as an extended trial period.

Every week or two, annotators were given a new batch of assignments in a new genre, beginning with popular, which was expected to be the least problematic, and continuing in order with jazz, classical, and world, which were predicted to be of increasing difficulty. At the end of the six weeks, supervision of the annotators was relaxed and any problems addressed on an *ad hoc* basis. Data collection continued over the next 12 weeks, by which point the majority of assignments had been completed.

We collected the self-reported time it took to produce each annotation in order to assess productivity. The times are plotted as a function of the date for the first 1700 annotations in Figure 2. It can be seen that, disregarding a number of outliers towards the beginning of the project, annotation time decreased modestly, from a mode of 20 minutes in the first 100 days, to a mode of 15 minutes in the remainder, enough for 3 full listenings of the average song, which was 4:21 long. The average annotation time also dropped from 21 to 17 minutes. Earlier analysis showed a slight correlation between a song's length and its annotation time.

3.4.1 Annotation format and procedure revisions

After each new assignment, we solicited feedback from the annotators on what weaknesses or ambiguities in the annotation format and procedure were revealed. Most issues were addressed and resolved at regular group meetings, where we also planned and agreed on the vocabulary. Feedback led to the introduction of new heuristics (e.g., we established a preference to have segment boundaries fall on downbeats, even in the presence of pickups). In one case, feedback led to a major revision of the format. We originally used the "V1" and "V2" markers described by [10] to implicitly encode musical similarity at a shorter timescale. However, annotators found that explicitly describing the structure at both timescales was both conceptually simpler and quicker. Annotators were satisfied by the switch and the subsequent annotations were also more meaningful.

4. RESULTS

In this section we report certain properties of the collected data, including inter-annotator agreement. From the variety of existing measures commonly used to compare two annotations (defined in [12] among others), we estimate the pairwise *f*-measure, boundary *f*-measure, and Rand index.

The average number of segments per annotation was 11.3 for the large-scale analyses, with half of the analyses having between 8 and 14 segments. These figures were 38.4 and between 20 and 49 for the small-scale analyses. On average, there were 4.0 unique large-scale labels and 7.2 unique small-scale labels per annotation.

In the evaluation of each, we consider one annotation as ground truth and the other as an estimate. Boundary *f*-measure is found by observing the precision and recall with which the estimated boundaries reproduce the ground truth boundaries. Boundaries are correct if they lie within some tolerance window (0.5 or 3 seconds) of a true boundary. Pairwise *f*-measure treats all pairs of frames with the same ground truth label as a set of similarity relationships that the estimated description retrieves with some precision and recall. The Rand index is similar except that it also identifies how many non-matching pairs of frames were correctly retrieved. The agreement between 974 pairs of annotations are reported in Table 3.

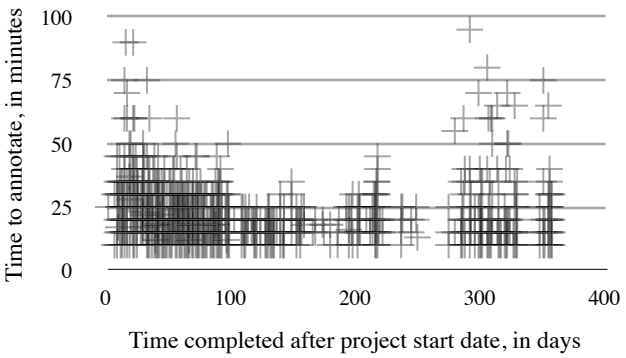

Figure 2. Plot of annotation times over the course of the project timeline.

Annotations compared	PW *f*	Rand index	Bound *f* (0.5 sec)	Bound *f* (3 sec)
1. Large-large	0.76	0.79	0.69	0.77
2. Small-small	0.69	0.81	0.73	0.82
3. Small-large and large-small (average)	0.60	0.70	0.38	0.44
4. Best case	0.81	0.87	0.80	0.89

Table 3. Average agreement between 974 pairs of annotations, as estimated by four similarity metrics (pairwise *f*-measure, Rand index, and boundary *f*-measure with two thresholds) when comparing: (1) both annotators' large-scale annotations; (2) both small-scale annotations; (3) one annotator's large-scale annotation and the other's small-scale one. The last row (4) takes the maximum similarity of all four possible pairings between the first and second annotators' musical similarity labels.

Each annotation describes musical similarity at two levels of detail, both of which should be considered valid descriptions. To compare two annotations, we may compare the large-scale labels only or the small-scale labels only, but we may also find the similarity of all pairs (including small-to-large and large-to-small) and take the maximum similarity to estimate the inter-annotator agreement. This will allow us to recognize cases where the annotators have focused on different timescales. As seen in Table 3, the agreement between large-scale labels (pairwise f = 0.76, Rand = 0.79) is comparable to that between small-scale

labels (pairwise $f = 0.69$, Rand= 0.81), and the average best match found is slightly higher than each (pairwise $f = 0.81$, Rand = 0.87). For comparison, [8] reported a pairwise f of 0.89 on a test set of 30 songs from the TUT set, and [1] reported a boundary f measure of 0.91 (using a 0.75-second threshold) on a test set of 20 songs.

The agreement was not found to depend greatly on the genre. This is reasonable since each of the broad genres considered here are each very diverse and contain some straightforward and some complex pieces. For instance, the popular genre includes both straightforward pop music and more difficult to annotate progressive rock; likewise, though much world music poses a challenge to annotators, subgenera such as klezmer and Celtic music can be structurally straightforward.

We replicated annotations for 97 recordings in the RWC data set. The RWC annotations distinguish similar and identical repetitions of sections by adding letters to function labels (e.g., "verse A", "verse B", etc.). We created two versions of the RWC labels, one retaining and one ignoring the additional letter labels. These were compared to the large- and small-scale SALAMI annotations, revealing modest agreement (see Table 4). Aside from the Rand index, the results indicate that the large-scale SALAMI analyses are more similar to the RWC annotations than the small-scale analyses.

Annotations compared	PW f	Rand index	Bound f (0.5 sec)	Bound f (3 sec)
1. RWC and large	0.64	0.73	0.57	0.75
2. RWC and small	0.45	0.77	0.38	0.52

Table 4. Average agreement between 97 pairs of RWC and SALAMI annotations when comparing: (1) SALAMI's large-scale labels with RWC's function class labels; (2) SALAMI's small-scale labels with RWC's distinct labels.

5. CONCLUSION

The SALAMI test set has over 2400 annotations describing the formal structure of almost 1400 pieces of music, from a wide variety of genres, including popular, jazz, classical, and world music. This set may be used for a variety of future studies: for example, on the connection between the surface characteristics of music and the perception of musical form, or between formal styles and musical parameters such as artist, genre, and place of origin. The test data and the hundreds of thousands of computed structural descriptions will soon be reachable from our website [11].

While the worth of the corpus will ultimately depend on the use researchers make of it, the quantity and richness of the information in the SALAMI test set should make it attractive to musicologists and music information retrieval researchers alike.

6. ACKNOWLEDGEMENTS

This research was funded by a Digging Into Data Challenge award, including funds from the National Science Foundation (under Grant No. IIS 10-42727), JISC, and the Social Sciences and Humanities Research Council of Canada. The authors would especially like to thank our annotators at McGill University and University of Southampton for their hard work.

7. REFERENCES

[1] Bimbot, F., O. Le Blouch, G. Sargent, and E. Vincent. 2010. Decomposition into autonomous and comparable blocks: A structural description of music pieces. *Proc. ISMIR*, 189–94.

[2] Bruderer, M., M. McKinney, and A. Kohlrausch. 2009. The perception of structural boundaries in melody lines of Western Popular music. *Musicæ Scientæ,* 8 (2): 272–313.

[3] Cannam, C., C. Landone, M. Sandler, and J. P. Bello. 2006. The Sonic Visualiser: A visualisation platform for semantic descriptors from musical signals. *Proc. ISMIR,* 324–7.

[4] Goto, M. 2006. AIST annotation for the RWC Music Database. *Proc. ISMIR,* 359–60.

[5] Isophonics datasets, Centre for Digital Music. http://www.isophonics.net/datasets.

[6] Live Music Archive. http://www.archive.org/details/etree.

[7] McKay, C., D. McEnnis, and I. Fujinaga. 2006. A large publicly accessible prototype audio database for music research. *Proc. ISMIR,* 160–3.

[8] Paulus, J., and A. Klapuri. 2009. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE TASLP*, 17 (6): 1159–70.

[9] Paulus, J., M. Müller, and A. Klapuri. 2010. Audio-based music structure analysis. *Proc. ISMIR*, 625–36.

[10] Peeters, G., and E. Deruty. 2009. Is music structure annotation multi-dimensional? A proposal for robust local music annotation. *Proc. LSAS,* 75–90.

[11] SALAMI. http://salami.music.mcgill.ca/.

[12] Smith, J. B. L. 2010. A comparison and evaluation of approaches to the automatic formal analysis of musical audio. MA thesis, McGill University.

[13] TUTstructure07 dataset, Technical University of Tampere. http://www.cs.tut.fi/sgn/arg/paulus/TUTstructure07_files.html.