AUTOMATIC INTERPRETATION OF MUSIC STRUCTURE ANALYSES: A VALIDATED TECHNIQUE FOR POST-HOC ESTIMATION OF THE RATIONALE FOR AN ANNOTATION

Jordan B. L. Smith
National Institute of Advanced Industrial Science and Technology (AIST), Japan
jordan.smith@aist.go.jp

Elaine Chew
Queen Mary University of London
elaine.chew@qmul.ac.uk

ABSTRACT

Annotations of musical structure usually provide a low level of detail: they include boundary locations and section labels, but do not indicate what makes the sections similar or distinct, or what changes in the music at each boundary. For those studying annotated corpora, it would be useful to know the rationale for each annotation, but collecting this information from listeners is burdensome and difficult. We propose a new algorithm for estimating which musical features formed the basis for each part of an annotation. To evaluate our approach, we use a synthetic dataset of music clips, all designed to have ambiguous structure, that was previously used and validated in a psychology experiment. We find that, compared to a previous optimization-based algorithm, our correlation-based approach is better able to predict the rationale for an analysis. Using the best version of our algorithm, we process examples from the SALAMI dataset and demonstrate how we can augment the structure annotation data with estimated rationales, inviting new ways to research and use the data.

1. INTRODUCTION

Listeners perceive structure in music, and trying to predict the structures they perceive is a popular task in the MIR community [14]. Since the perception of structure is a complex phenomenon, the community focuses on a simpler, operational version: we imagine that structure, as perceived, can be characterized as a set of time points regarded as boundaries, and a set of labels that indicate which of the intervening segments repeat similar material. This simplification is not made naïvely: those who create annotations of musical structure are aware of its limitations, and the methodologies for annotating [1, 16, 21] and evaluating [7, 9, 11] structural analyses have become their own important subtopics in MIR.

Still, the simplification is unfortunate because musical similarity is multi-dimensional. If a listener declares that two excerpts are "similar", they could mean with respect to melody, contour, rhythm, timbre, or any combination of these or other musical attributes. This is in addition to the issue that structure itself is multi-dimensional; as pointed out in [16], boundaries may be perceived for reasons of musical similarity, musical function, or instrumentation.

Thus, in the transition from structure perception to structure annotation, we usually fail to capture *why* a listener has included a boundary or chosen a label. This information, if preserved (or reconstructed), would help us to understand the content of the annotations, and could lead to fairer evaluations of structure segmentation algorithms. It would also provide more meaningful data to analyze in musicology or music perception research.

How feasible is it to collect this information? As we found in [23], to transcribe the rationale for every aspect of an annotation is difficult and requires prolonged self-interrogation. Even before that, it is difficult to decide what information to collect, and how to collect it: should the data be collected after a listener has provided the segmentation, in the manner of music perception experiments [2]? Or should each piece be annotated several times, each time with a focus on a single feature [19]? No matter how it is done, collecting this information is burdensome.

A more practical possibility is to estimate this information automatically from existing annotations, which was our motivation in [22]. Our algorithm compared self-distance matrices (SDMs) for different features to the ground truth annotation, and found which parts of the feature-based SDMs best re-created the annotation-based SDM. While [22] presented some examples to demonstrate the plausibility of the approach, we offered no experimental validation.

Validation requires paired responses: a set of listeners' analyses, and the listeners' justifications for each analysis. Producing this data is time-consuming and burdensome for the reasons described above. However, we recently produced data suited to this purpose for a music perception study [20]. The goal of that study was to determine what role attention plays in the perception of structure.

In this article, we make three main contributions: first, we test whether the approach described in [22] can effectively predict the attention of the listeners, based on the dataset created for [20]. Second, we explain some short-

comings of the previous approach, and suggest and test two improvements. Third, we demonstrate how the validated algorithm can be used to analyze and to augment real-world data with new information layers.

The next two sections recap the studies on which this article builds. In Section 2, we briefly recall [22]'s algorithm, point out some shortcomings, and introduce a refined approach. In Section 3, we summarize the results of the experiment in [20], and describe in more detail the data developed for that study and used in this one. In Sections 4 and 5, we outline the validation experiment and discuss the results, and in Section 6 we use the algorithm to create new information layers for examples from SALAMI [21]. We close with a few observations on the limitations of the present work and recommendations for future research.

2. AN ALGORITHM FOR ESTIMATING FEATURE RELEVANCE

In [22] we estimated the relevance of musical features to a listener's analysis section-by-section by finding the weighted sum of feature-derived SDMs that best matched the analysis. The analysis is represented as a binary SDM, expanded to the same timescale as the feature SDMs. A number n of feature matrices are computed; from each, we derive m single-section SDMs by taking only the rows and columns associated with that single segment, as defined by the annotation. (This row and column selection is done by multiplying the SDM with a segment mask.) This gives $n \cdot m$ component matrices. A quadratic program (QP) is used to find the weights for these components whose sum optimally reproduces the annotation-derived SDM; these weights, the reconstruction coefficients, are taken to indicate feature relevance.

The method is illustrated in Fig. 1. The sound example has the form ABAB with respect to harmony, AABB with respect to rhythm, and ABBA with respect to timbre. If a listener gives the analysis ABAB, segmenting the audio at the 1/4, 2/4 and 3/4 marks, we obtain the four segment masks given in the top row. We compute four audio features, each related to a different musical attribute (see Section 4.1 for details), which are pointwise multiplied by the masks to give 8 potential components. The QP finds the optimal combination of components to reproduce the annotation in the top-left corner, and gives the coefficients shown above each component. In this case, the algorithm has successfully identified that bass chroma is the feature that best justifies the analysis.

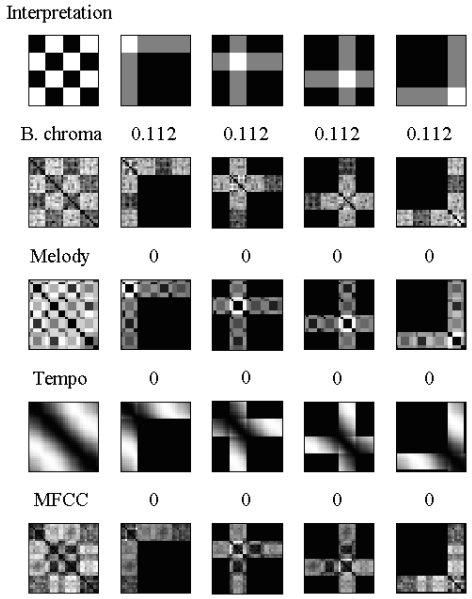

Figure 1. Illustration of component-building for QP algorithm. Four beat-indexed feature matrices (at left) are multiplied by the masks (top) given by the segmentation, which here is ABAB. The number above each component is the QP's estimate of the component's importance.

2.1 Algorithm Improvements

One limitation of this approach is that none of the feature matrices may properly reflect the homogeneity of a given section. We could include additional SDMs that have been smoothed at different timescales (as demonstrated in [22]), but the smoothing can blur the boundaries between sections even as they make the sections more homogeneous. We could use stripe-based instead of block-based masks in order to capture repetitions of feature sequences, but in non-square blocks (which occur whenever two segments have unequal lengths), it is not easy to guess the best orientation or placement of the stripes.

A second problem is that it is unclear how to interpret some aspects of the QP. Should the individual reconstruction coefficients, or their sum, be bounded? Leaving them unbounded can lead to unconstrained solutions, but if bounds are imposed, how should they be interpreted?

A third problem is that by finding the single optimal *sum* of matrix components, some good explanations may be ignored. For example, if there are two matrix components which both justify a particular part of the analysis, the QP may find that only one is necessary. Thus, we cannot conclude that features omitted from the solution are necessarily irrelevant, which is a big limitation.

For the first problem, we propose that instead of using the original SDMs, with all their heterogeneities, we reduce them to segment-indexed SDMs, a common practice since [4]. Similar to [13], we may take the distance between each pair of segments to be the average distance of all the pixels in the submatrix over which the segments intersect. The segment-indexed SDM can then be analyzed with the QP as before, although with a substantial reduction in complexity.

A second way to address the problem is to use a diagonal stripe-based mask instead of a block-based mask. Since the diagonals are the most salient portions of the SDM, it makes sense to focus on reconstructing this portion of the SDM. Emphasizing stripes is a common SDM analysis technique, and a comparison of block and stripe features found that when boundaries were given, stripe fea-

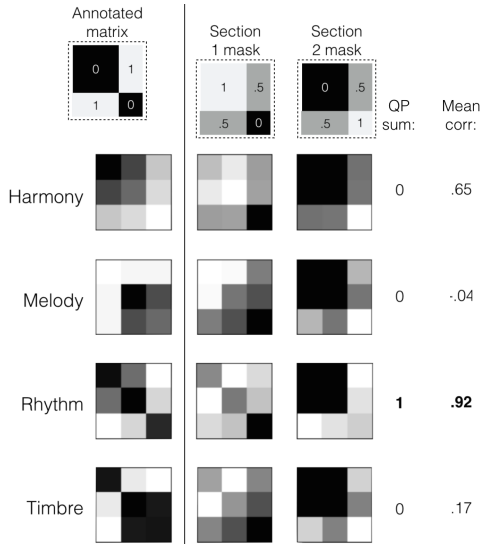

Figure 2. Illustration of proposed segment-indexed approach, with both QP- and correlation-based estimates of feature relevance.

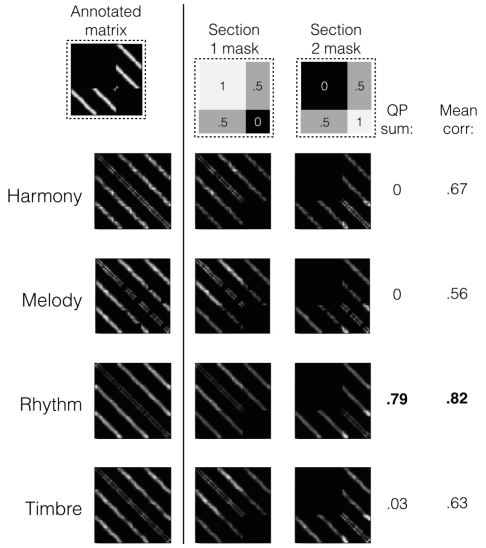

Figure 3. Illustration of proposed stripe approach with correlation. Like in Fig. 2, QP coefficients are in the middle column, correlations on the right.

tures were more effective [12]. We remember the caveat above, that repeated segments with different durations pose a problem for creating a stripe-based mask, but we can still test it in cases where this is not an issue.

A third proposal, which addresses the second and third problems above, is to dispense with the QP altogether and simply take the element-wise Pearson correlation between the feature-derived and annotation-derived matrices. (The same section-by-section method still applies.) Correlations are perhaps more intuitive than QPs and reconstruction coefficients, and using them would permit second-place features to be more readily identified in the solution.

Fig. 2 shows the output for an example of a three-part stimulus, using the suggested improvement of segment-indexing: the features have been averaged over the blocks given by the segmentation. The sum of the reconstruction coefficients obtained using the QP method are given in the "QP sum" column, and the mean point-wise correlation between the masked regions is shown in the "Mean corr." column. Fig. 3 shows the output for the same example but using the stripe-based mask. The mask is constructed by drawing a diagonal line across each block of the original beat-indexed SDM, and then applying a 2D convolution with a Gaussian kernel of width 5 beats.

To sum up, we suggest three improvements to the algorithm: (1) using the correlation between submatrices, instead of QP, to estimate their relevance; (2) using a segment-indexed version of the SDM; and (3) applying a stripe mask to the SDM, instead of using the blocks.

3. A DATASET OF VALIDATED ANALYSES

Researchers in music psychology, like those in MIR, are invested in modeling how listeners perceive structure. (For one discussion, see [15].) The goal of [20] was to determine whether listeners could be influenced to perceive different structures by manipulating the musical feature to which they paid attention. In order to test this, we composed a set of artificial musical stimuli in which four different features (harmony, melody, rhythm and timbre) were systematically changed at different times, creating musical passages with ambiguous forms. These four features were chosen because they figured most prominently in studies where listeners were asked to justify why they perceived a given boundary, such as [2].

The three-part stimuli had two potential structures, AAB or ABB, with different features changing at different times. For example, the passage in Fig. 4a has form AAB with respect to harmony, and form ABB with respect to melody. The four-part stimuli had three potential structures, AABB, ABAB or ABBA, so that at every boundary there were two features that changed. For example, in the passage in Fig. 4d, the rhythm and harmony both change after the second measure.

As stated above, validating the algorithm requires musical examples where listeners' analyses are paired with their justifications—i.e., with the musical attributes to which they were paying attention. Many datasets of structural analyses exist, but none indicate which musical attributes justify the analyses. Also, in typical pieces of music, attributes change frequently, to different extents, and often simultaneously. To validate this algorithm we should use music with known, controlled changes. Hence, artificial stimuli such as these are valuable resources to validate the algorithm: each passage contains precise change points related to known musical attributes; and the link between the attributes and the different forms has been affirmed by listeners in an experimental setting.

More artificial stimuli could be generated and tested in future work; this may be a convenient way to provide deep-learning algorithms with the quantity of labelled data

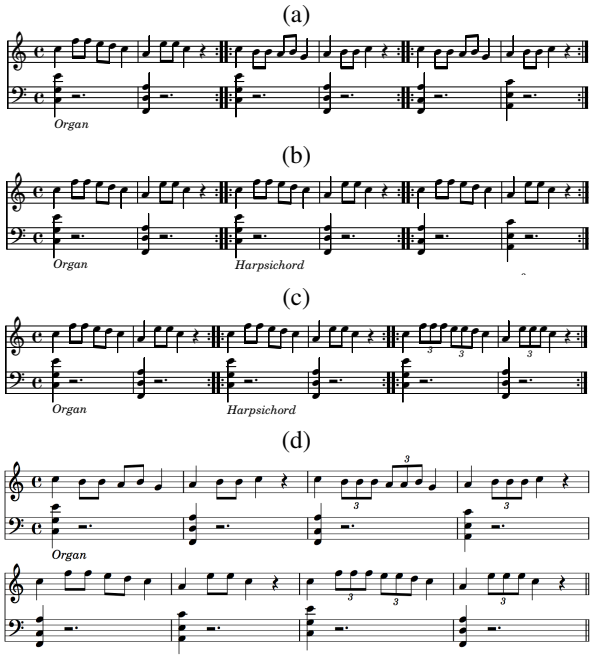

Figure 4. (a) Example stimulus with harmonic form AAB and melodic form ABB. (b) Harmonic form AAB, timbral form ABB. (c) Rhythmic form AAB, timbral form ABB. (d) Example four-part stimulis with melodic form AABB, rhythmic form ABAB, and harmonic form ABBA.

they require. However, it is not as simple as sonifying a symbolic score, since scores must be annotated in order to know the perceived structure and the musical features that motivate that analysis. The stimuli in our study are rare in that they were (1) composed so that musical features varied systematically, and (2) used in a listening experiment to validate that the intended structures were perceived, for the intended reason.

3.1 Stimulus Details

For [20], we composed three sets of stimuli. Each stimulus contains two voices, and in each set of stimuli, each voice potentially expresses changes in two different features. The examples in Fig. 4 are from the "HT-MR" set, where one voice expresses changes in harmony and timbre, and the other, changes in melody and rhythm. "HM-RT" and "HR-MT" sets were also composed.

Since in each set of stimuli, certain features are "convolved," some incorrect answers are less wrong than others. For instance, the feature that changes at the second boundary of the example in Fig. 4c is rhythm, but if an algorithm said that the boundary was justified by melody, it would be partially right.

The stems for the stimuli were composed using a Digital Audio Workstation with standard instrument patches. The 8 stems for each set were systematically recombined to generate 192 three-part stimuli and 384 four-part stimuli, for a total of 1728 stimuli among all sets. Efforts were made to keep constant all musical features other than harmony, melody, rhythm and timbre: the tempo of all stimuli is 140 bpm, and the loudness of each voice and each passage is approximately equal. The stimuli are now freely available on Github. [1]

4. EXPERIMENT

4.1 Features

The stimuli manipulated four different musical attributes (in three environments): harmony, melody, rhythm and timbre. We want to extract audio features that match each of these attributes independently. Each audio feature should change when the related musical feature changes, and be robust to changes in other musical features. We selected two audio features for each musical feature, all available as Vamp plugins [2] and listed in Tab. 1. We used ground truth beat locations, and median feature values were taken for each beat. Each dimension was normalized (independently for each stimulus) to zero mean, unit variance. All features were extracted using Sonic Annotator [3] using the default settings. For some features, we performed additional processing:

Chords: Chord labels were estimated from Chordino and reconverted back to a chroma-like representation. This feature is thus based on the same information as *bass chroma*, but refined with the chord-estimation algorithm.

Melody: The chroma of the estimated melody, and the interval between the current steady-state note and the previous one, each a 12-dimensional feature per frame. We also used the register of the melody: low, middle or high.

Autocorrelation: this was computed on an onset detection function with a sliding window.

Low level features: a concatenation of loudness, RMS amplitude, rolloff, sharpness, smoothness, tristimulus, zero-crossing rate, and the centroid, kurtosis, skewness, and slope of the spectrum.

Feature	***Vamp plugins used to obtain feature***
Harmony	*Bass chroma*, from Chordino and NNLS Chroma plugin [8] *Chord notes* [8]
Melody	*Treble chroma* [8] *Melody*, based on MELODIA [18]
Rhythm	*Cyclic tempogram* [6] *Autocorrelation*, based on UAPlugins's Note Onset Detector [17]
Timbre	*MFCCs* (2nd to 13th), from Chris Cannam and Jamie Bullock's LibXtract library *Low level features*, a set of fifteen one-dimensional descriptors from LibXtract

Table 1. List of features chosen, and Vamp plugins used to obtain them

[1] https://github.com/jblsmith/music-structure-stimuli.

[2] vamp-plugins.org

4.2 Results

We applied the algorithms, discussed in Section 2, on the stimuli discussed in Section 3. For each three-part stimulus, we ran the algorithms twice: once with analysis AAB, once with ABB. Likewise, we ran the algorithm thrice on each four-part stimulus to find the best justifications for forms AABB, ABAB and ABBA. Each algorithm takes one of these analyses as input. The output of each algorithm is a matrix of feature relevance values $x_{i,j}$: one per section i, per feature j. The importance of feature j is the sum across all the sections: $s_j = \sum_i x_{i,j}$. The importance of each musical attribute a is the sum of the two values s_j related to that feature: $y_a = s_{a1} + s_{a2}$. We end up with four values y_a.

We test whether the maximum value correctly predicts the feature related to the analysis with $\arg\max_a y_a$. The fraction of trials with correct guesses is the accuracy. Each trial has one focal pattern and three potential wrong answers, so the random baseline performance is 25%.

The five algorithm options were: whether to use cosine or Euclidean distance (in either case, the values were rescaled between 0 and 1); whether to compute beat-indexed or segment-indexed SDMs; whether to apply stripe-based masks to the SDMs; whether to use the QP or correlation-based approach; and finally, if using QP, what constraint to use. We tested three constraints: (a) $\sum_{i,j} x_{i,j} = C$ (the sum of the coefficients over the entire piece has a fixed value); (b) $\sum_j x_{i,j} = C$ (the sum of the coefficients for each section in the piece has a fixed value); and (c) $0 \leq x_{i,j} \leq 1$. These options were tested in a full-factorial design, replicated across three variables that were not part of the algorithm: the relevant feature; the music environment; and the stimulus length (3 or 4 sections).

We fit a linear model to the results and used ANOVA to interpret the eight factors. With 268,512 trials, three factors were insignificant ($p > 0.05$): stimulus length, distance metric, and QP constraint. The other five factors all had $p < 0.0001$, and main effect plots for each are shown in Fig. 5. They show that performance varied greatly among the music examples and features. However, the three proposed changes to the original algorithm—using correlation instead of QP, using stripe masks, and using segment-indexed SDMs—all saw improvements, albeit a minor one in the case of segment indexing.

Tab. 2 gives the accuracy for different parameter settings. It shows that although the main effects appear modest in Fig. 5, their impact is additive: the original approach achieved 47% accuracy, and the three changes (using correlation, segment-indexing, and applying a stripe mask) together raised the accuracy to nearly 70%.

These are the accuracies for choosing the most correct answer, but not all errors are equally bad: guessing a feature that was convolved in the stimulus with the correct one is sometimes a fair mistake. However, Tab. 2 shows that the "convolved-with-correct" answer was not given any special weight by the algorithms. There are 3 features besides the correct one, so the chance of randomly guessing the convolved feature is 33%. In all cases, fewer than a third

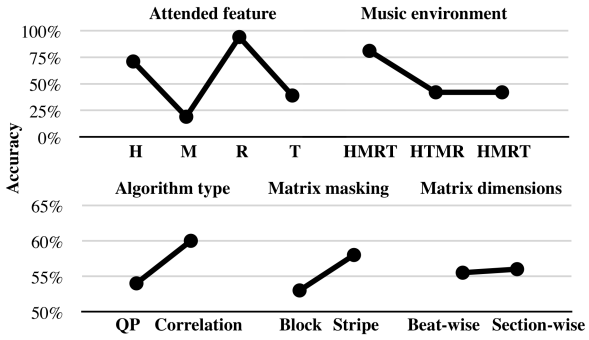

Figure 5. Main effect of significant factors on accuracy (i.e., rate of correct guesses).

Method:	**Quad. Prog.**		**Correlation**	
Settings:	Correct	Conv.	Correct	Conv.
Regular	47.1	13.7	52.3	12.8
Seg.-indexing	46.9	16.3	60.6	8.9
Stripe mask	52.4	13.5	62.4	11.9
Seg. and stripes	59.6	12.8	69.6	7.2

Table 2. Comparison of QP-based and correlation-based algorithms. Columns indicate how often the guessed feature was correct ("Correct") or convolved with the correct feature ("Conv."). For example, in the HT-MR environment, if the correct feature for a trial is timbre, guessing harmony could be half-right.

of the incorrect answers related to the convolved feature.

Prediction accuracy varied greatly among the features, as can be seen in the confusion matrices for the algorithms. Three are shown in Fig. 6, one for each music environment. These are the results for the best-performing algorithm. For harmony, we can observe that chord notes were more effective than bass chroma, the feature from which they derive. Bass chroma were especially misled in the HM-RT setting, possibly due to the difference in bass drum between the two timbre settings. With melody, it was also the case that the 2nd-order feature (the estimated predominant pitch and interval) was better than the lower-level feature (treble chroma).

5. DISCUSSION

The results validate the algorithm proposed in [22]. However, they also show that a simpler correlation-based approach is better at predicting how best to justify an analysis: it outperformed the QP approach by roughly 10%. Two other refinements, the stripe-based mask and the segment indexing, increased accuracy by roughly another 10%.

However, the confusion matrices revealed great disparities between the features we chose to use: some, such as Chordino, were effective; others, such as the tempogram and MFCCs, were often wrong. Arguably, it is naïve for us to presume that off-the-shelf features can detect the types of musical changes we created in the stimuli. Perhaps it is no accident that the four features we tweaked or assem-

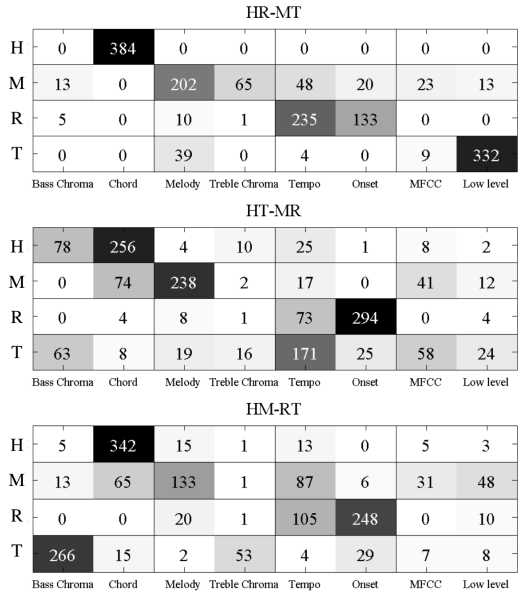

Figure 6. Confusion matrix for algorithm using correlation, stripe masks, and segment-indexed SDMs. The rows gives the correct musical attribute; the column indicates the audio feature with maximum relevance.

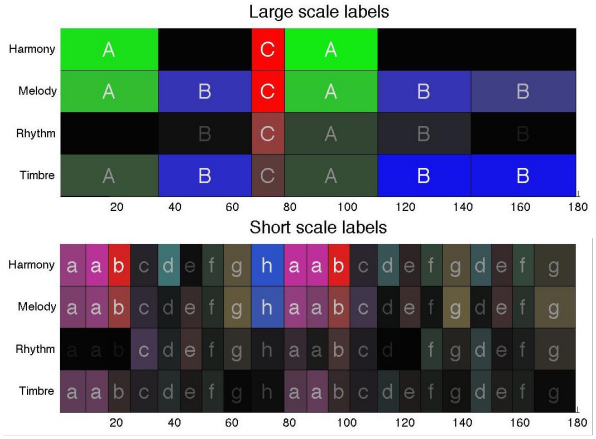

Figure 7. Example augmented annotation for the song "We Are The Champions" by Queen. The letters and colors both encode the section labels. Brightness indicates a feature's relevance to a section.

bled for this purpose (chord notes, MELODIA-based feature, onset autocorrelation and low-level features) tended to outperform their off-the-shelf rivals.

Still, the underperformance is surprising, since the stimuli are highly constrained: in the study for which the stimuli were created, listeners identified the attribute that changed at a boundary with 85% accuracy [20]. It seems reasonable to expect that, say, MFCCs will change more when a trumpet is swapped for a flute than when a trumpet plays a different melody; or that when the harmony changes, bass chroma will be more affected than the tempogram. Yet these are among the errors made by the features in this study. The results thus remind us of the utility of carefully-designed features, such as timbre-invariant chroma [10].

An alternative to testing hand-crafted features is to learn features with deep learning, but as mentioned earlier, this would require building a much larger, more representative stimulus set—more stimuli than can easily be validated in a listener study. The small set used here is suitable for testing existing features, but not learning new ones.

6. APPLICATION

The correlation algorithm can be used to interpret annotations in the SALAMI corpus [21]. We used the segment-indexing setting but not the stripe-masking, which (as noted in Section 2.1) is not applicable when unequal segment lengths give rectangular blocks. The audio processing was the same except that BeatRoot [5] was used to locate beats.

Fig. 7 visualizes a listener's analysis of "We Are The Champions" by Queen at the long and short timescales. Each vertical slice corresponds to a single section, and the brightness of each cell indicates the correlation of that feature to that section. We can see that on a long timescale, the verse sections (A) were characterized by their harmonic and melodic content, while the chorus sections (B) were characterized more by their timbre. However, on a short timescale, subsection a was also characterized by timbre, and many of the subsections of B were more strongly characterized by harmony and melody compared to B itself.

This, it turns out, is an accurate description of the song: in a, Freddie Mercury sings above a piano and bass only; the electric guitar enters quietly in b, but the drums come in with c in a raucous crescendo to the chorus. The timbral inconsistency of A means that timbre would be a poor feature to use to justify grouping the first four subsections into a larger unit.

On the other hand, the timbre of the choruses is relatively homogeneous; this makes it a good feature to justify grouping the B sections together, but also makes it a poor feature to justify giving the subsections of B different labels. The fact that subsections d, e, f and g have different labels must therefore reflect their pitch content.

7. CONCLUSION

We have validated the algorithm proposed by [22], and proposed three modifications to improve its effectiveness. Although we restricted this study to stimuli that were validated in a psychology experiment, it would be possible to generate large amounts of artificial music, with more complicated patterns of repetition and variation, and changes in more musical parameters, like loudness, tempo, syncopation, dissonance, and so on.

The accuracy of the algorithms fell short of human performance. Given the disparities among the features, this must be due in part to the mismatch between the audio features we chose and the musical attributes manipulated in the stimuli. Despite this, the algorithm is useful for visualizing the structure of pieces in a new way: by highlighting the musical features that explain the annotation.

8. REFERENCES

[1] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent. Semiotic structure labeling of music pieces: Concepts, methods and annotation conventions. In *Proceedings of ISMIR*, pages 235–240, Porto, Portugal, 2012.

[2] Michael Bruderer, Martin McKinney, and Armin Kohlrausch. The perception of structural boundaries in melody lines of Western popular music. *Musicæ-Scientæ*, 13(2):273–313, 2009.

[3] Chris Cannam, Michael O. Jewell, Christophe Rhodes, Mark Sandler, and Mark d'Inverno. Linked data and you: Bringing music research software into the semantic web. *Journal of New Music Research*, 39(4):313–325, 2010.

[4] Matthew Cooper and Jonathan Foote. Summarizing popular music via structural similarity analysis. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 127–30, New Paltz, NY, United States, 2003.

[5] Simon Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

[6] Peter Grosche, Meinard Müller, and Frank Kurth. Cyclic tempogram - a mid-level tempo representation for music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, 2010.

[7] Hanna Lukashevich. Towards quantitative measures of evaluating song segmentation. In *Proceedings of ISMIR*, pages 375–380, Philadelphia, PA, USA, 2008.

[8] Matthias Mauch and Simon Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proceedings of ISMIR*, pages 135–140, Utrecht, Netherlands, 2010.

[9] Brian McFee, Oriol Nieto, and Juan Pablo Bello. Hierarchical evaluation of segment boundary detection. In *Proceedings of ISMIR*, Málaga, Spain, 2015.

[10] 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, 2010.

[11] Oriol Nieto, Morwaread Farbood, Tristan Jehan, and Juan Pablo Bello. Perceptual analysis of the f-measure to evaluate section boundaries in music. In *Proceedings of ISMIR*, Taipei, Taiwan, 2014.

[12] Jouni Paulus and Anssi Klapuri. Acoustic features for music piece structure analysis. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 309–312, Espoo, Finland, 2008.

[13] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–1170, 2009.

[14] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of ISMIR*, pages 625–636, Utrecht, The Netherlands, 2010.

[15] Marcus T. Pearce, Daniel Müllensiefen, and Geraint A. Wiggins. The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, 39:1367–1391, 2010.

[16] Geoffroy Peeters and Emmanuel Deruty. Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proceedings of the International Workshop on Learning the Semantics of Audio Signals*, pages 75–90, Graz, Austria, 2009.

[17] Antonio Pertusa and José Manuel Iñesta. Note onset detection using one semitone filter-bank for mirex 2009. In *MIREX Audio Onset Detection*, Kobe, Japan, 2009.

[18] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, 20:1759–1770, 2012.

[19] Chris Sanden, Chad R. Befus, and John Z. Zhang. A perceptual study on music segmentation and genre classification. *Journal of New Music Research*, 41(3):277–293, 2012.

[20] Jordan B. L. Smith. Explaining listener differences in the perception of musical structure. September 2014.

[21] Jordan B. L. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of ISMIR*, pages 555–560, Miami, FL, United States, 2011.

[22] Jordan B. L. Smith and Elaine Chew. Using Quadratic Programming to estimate feature relevance in structural analyses of music. In *Proceedings of the ACM International Conference on Multimedia*, pages 113–122, Barcelona, Spain, 2013.

[23] Jordan B. L. Smith, Isaac Schankler, and Elaine Chew. Listening as a creative act: Meaningful differences in structural annotations of improvised performances. *Music Theory Online*, 20(3), 2014.