Using Quadratic Programming to Estimate Feature Relevance in Structural Analyses of Music

Jordan B. L. Smith
Centre for Digital Music
Queen Mary University of London
Mile End Road, London, UK
jordan.smith@eecs.qmul.ac.uk

Elaine Chew
Centre for Digital Music
Queen Mary University of London
Mile End Road, London, UK
elaine.chew@eecs.qmul.ac.uk

ABSTRACT

To identify repeated patterns and contrasting sections in music, it is common to use self-similarity matrices (SSMs) to visualize and estimate structure. We introduce a novel application for SSMs derived from audio recordings: using them to learn about the potential reasoning behind a listener's annotation. We use SSMs generated by musically-motivated audio features at various timescales to represent contributions to a structural annotation. Since a listener's attention can shift among musical features (e.g., rhythm, timbre, and harmony) throughout a piece, we further break down the SSMs into section-wise components and use quadratic programming (QP) to minimize the distance between a linear sum of these components and the annotated description. We posit that the optimal section-wise weights on the feature components may indicate the features to which a listener attended when annotating a piece, and thus may help us to understand why two listeners disagreed about a piece's structure. We discuss some examples that substantiate the claim that feature relevance varies throughout a piece, using our method to investigate differences between listeners' interpretations, and lastly propose some variations on our method.

Categories and Subject Descriptors

H.5.5 [Sound and Music Computing]: Methodologies and Techniques.

General Terms

Algorithms, Measurement, Experimentation, Human Factors, Theory.

Keywords

Music structure analysis, music perception, music cognition, repetition, quadratic programming.

1. INTRODUCTION

One of the most important aspects of music is that it is repetitive: individual sounds, notes and chords, dynamic gestures, rhythmic patterns, instrumentations, and so forth are all elements liable to repeat, whether identically or with some variation. The structure of repeating elements, while an abstract concept, is itself one of the most salient aspects of music to a listener. In a formal analysis, a listener can indicate which sections they believe are repetitions, and many of those studying music theory and music perception are interested in discovering how listeners make the analysis decisions they do. The task of discovering repetitions is also important in music information retrieval (MIR), since many MIR tasks can be performed or improved upon using information about music structure: for example, cover song detection [12] and chord transcription [19].

Recurrence plots, proposed by [4] for analyzing the motion of dynamical systems, can reveal repetitions in sequential data. A self-similarity matrix (SSM), originally proposed by [7], is a variation on such plots that moves beyond the usual feature-based methods of visualizing music, such as pitch rolls and other time-frequency representations. Unlike these, SSMs do not represent the musical content itself, but only the pattern of repetitions and recurrences that it contains. The features themselves are of course still important in any given application: an SSM based on pitch content may be effective for tracking melodic repetitions but not repetitions of percussive sounds. In MIR, SSMs have proven highly useful for estimating the large-scale structure that listeners find salient (for a review, see [27]).

In this article, we argue that SSMs could also be used for the inverse problem: studying the annotations of listeners to discover what aspects of the music they found salient. In particular, we show how the multiple SSMs generated from different features, as well as the segmentation of the listener, can be exploited to model which features were more important at which point in the piece. The approach is premised on the assumption that the structure of a piece of music can be represented as the weighted sum of feature matrix components, and that by finding the weights of these feature components, we in essence can model the potential reasons a listener might give to justify the analysis.

Accounting for listener attention is a key aspect of user modelling, for example, in music recommendation systems and user-centered music interfaces. The emphasis in this article is on motivating and defining a novel methodology, supported by several real-world examples. The approach has the potential for extensions to other multimedia domains—for example, to relate video annotations with time series data.

1.1 Previous methods in SSM calculation

Whereas earlier experiments with extracting structure based on SSMs focused on one feature at a time (e.g., [7] made SSMs de-

MM '13, October 21 - 25 2013, Barcelona, Spain

ACM 978-1-4503-2404-5/13/10…$15.00.
http://dx.doi.org/10.1145/2502081.2502124

rived only from Mel-frequency cepstral coefficients (MFCCs), and [1] used only chroma for chorus detection), it was soon realized that using multiple features could improve results. Eronen [6] calculated SSMs from MFCCs and chroma and summed the result, while [18] obtained three SSMs from chroma vectors, each calculated to reflect repetitions at different time scales, and took the element-wise product of the trio to reduce noise. Paulus & Klapuri's [26] optimization-based approach used information from separate SSMs reflecting timbral, harmonic and rhythmic similarity. In order to find transposed repetitions, [11] searched for maxima across multiple chroma-based SSMs. Rather than generate separate SSMs for separate features, [13] concatenated the features vectors for each frame and calculated a single SSM from the result.

However, simple combinations of SSMs rarely result in the exact structure that the experimenter hopes to extract. While SSMs provide an intuitive visualization of a piece, successful methods of automatic music analysis based on SSMs invariably employ complex post-processing steps to obtain the answer: this has taken the form of low- and high-pass filtering [11], erosion/dilation operations [21], dynamic programming [30], non-negative matrix factorization (NMF) [16], re-emphasis of transitive relationships [29], and so on. We may conclude that a simple sum of SSMs does not reflect the similarity judgements that a listener may make across an entire piece.

What information could the sum of SSMs be missing? One straightforward suggestion is a weight for each feature. Perceptual evidence supports this strategy: for example, in investigating the relative importance of different laws in a Generative Theory of Tonal Music, [10] found that laws relating to some features are more important than others. The idea to tune feature weights before analyzing structure has appeared before: to improve a structural segmentation algorithm, [25] chose feature weights to maximize the separability of vectors according to the Fisher criterion. In [15], the size of the window for calculating features was adapted to the estimated rate of change, improving the clarity of block patterns in SSMs. A hierarchical SSM proposed by [14] used different features and techniques at each time scale in a musicologically-informed manner.

Another aspect of listening that may be missed when SSMs are combined is that the focus of a listener's attention may shift at various points throughout a piece. For example, the self-similarity of a chorus of a given song may be very well accounted for by an SSM based on harmony, whereas the self-similarity of the guitar solo that follows may not be. Again, listener studies such as [2] and [3] demonstrate that the justifications listeners give for perceiving section boundaries can vary throughout a piece.

In the next section, we present a method of combining SSMs to exploit these facts of music perception—namely, that musical features vary in their relative salience, and that the attention of the listener is drawn to different features throughout a piece. Unlike previous work on adaptive parameter setting in music structure analysis [14, 15, 25], our aim is not to predict structure, but to discover connections between an audio recording and an accompanying analysis. Briefly put, our approach breaks a set of SSMs derived from various features into submatrix components, and then finds the optimal set of components to reconstruct the description of a piece given by a listener. The method may help those studying the cognition of musical form by illustrating what features best explain each part of the analysis. It may also be used to suggest reasons that two listeners disagree about the structure of a piece, an application discussed in Section 3.

2. PROPOSED METHOD

We first review how to calculate an SSM from acoustic data or from an annotation. In Section 2.2 we motivate our approach and explain the mathematics using a simple example, and in Section 2.3 demonstrate its use on the song "Yellow Submarine" by The Beatles.

2.1 Self-similarity matrix calculation

A self-similarity matrix (SSM), proposed by [7], can be thought of as a real-valued recurrence plot where element e_{ij} indicates the similarity between frame i and frame j of a sequence of frames. It is typical to use Euclidean or cosine distance [27] as a distance metric; here, we use cosine distance for its natural scaling between -1 and 1. Repeated sequences in recurrence plots are revealed as diagonal lines. In SSMs based on music, it is also common to see off-diagonal blocks, revealing the repetition of sections that are homogenous with respect to a given feature.

2.1.1 SSMs from annotations

Binary SSMs are commonly generated from structural annotations as diagrams (e.g., [27]) or to illustrate examples of song structure. Using cosine similarity, we set e_{ij} = 1 if frames i and j belong to sections with the same label, and -1 otherwise (see example in Figure 1). This section describes some variations on the usual approach that is relevant for our data.

The annotation in Figure 1, like all the examples in this article, are drawn from the Structural Analysis of Large Amounts of Music Information (SALAMI) dataset [31]. In the SALAMI annotation format, information about repeating sections is given at large and a small time scale, and sections may be distinguished with prime symbols (e.g., A vs. A'), which fuzzily indicates similarity with variation.

We include some of the richness of this description in the SSM by generating a separate SSM for each timescale and summing the

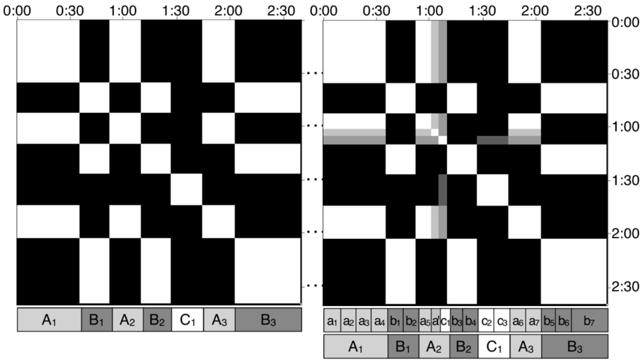

Figure 1. Left: SSM derived from annotation of "Yellow Submarine." Time progresses from left to right and from top to bottom. The large-scale annotation below it is from the SALAMI database (salami_id: 1634). Right: An alternative SSM derived using an additional layer of the annotation, where α, the relative weight of the large-scale labels, is set at 0.625 and β, the fractional similarity implied by primes, is 0.35.

results (see Figure 1). To emphasize one time scale over another, we can choose a weighting parameter $0 < \alpha < 1$, and multiply the large- and small-scale SSMs by α and $1 - \alpha$, respectively, before summing. A similar approach can act on prime symbols: when two frames have the same label but differ by a prime, instead of setting $e_{ij} = 1$, we can set it to some other value $-1 < \beta < 1$. Setting $\beta = 1$ would imply that A and A' are identical; $\beta = -1$ would imply they are completely distinct; and $\beta = 0$ would ignore the symbol.

2.1.2 SSMs from audio

The five different SSMs in Figure 2 were all calculated from a recording of "Yellow Submarine". Each one represents a different musical parameter: timbre, pitch, rhythm, tempo, and loudness.

MFCCs derive from the shape of a rescaled spectrum and hence can characterize a sound's timbre; chroma vectors estimate the power of each pitch class and hence characterize the harmonic content of the audio; fluctuation patterns (FPs) estimate the strength of low-frequency periodicities within Bark-scale frequency bands over windows that are several seconds long and hence characterize the rhythmic content [24]; periodicity histograms reflect the relative strength of different tempi by looking at sharp attacks in the audio and measuring the strength of periodicities in the tempo range of 40 to 240 bpm (0.6 to 4 Hz) [23]; and finally, the root mean square (RMS) of the waveform and the derivative of RMS estimate loudness and dynamic variations.

MFCCs and chroma were calculated using 0.19- and 0.10-second windows, respectively, with 50% overlap, using Queen Mary's Vamp Plugin set [17]. The twelve lowest MFCCs were kept, aside from the first, which correlates with loudness. FPs and periodicity histograms were calculated using 3-second windows and 0.37-second hops with the MA Toolbox [22]. FP vectors have 1200 elements, measuring 60 modulation frequencies in 20 Bark-scale frequency bands, while periodicity histograms have 2000 elements, indicating whenever any of 40 tempo ranges is activated beyond 50 fixed thresholds. While it is common to use dimensional reduction techniques to reduce the large size of the feature vectors, the relative differences between the raw vectors are still well captured in the SSMs in Figure 2. Lastly, RMS was calculated using 0.1-second windows and 50% overlap.

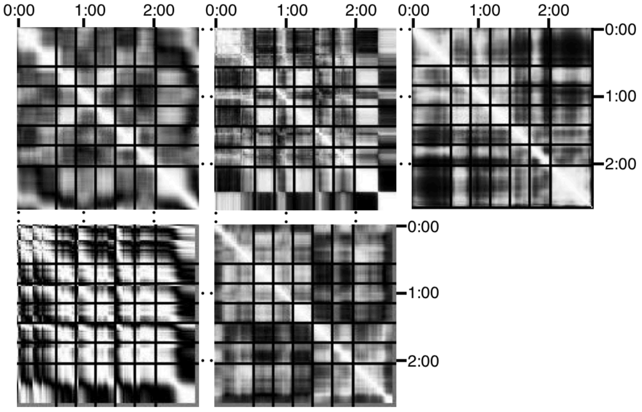

Figure 2. Five SSMs calculated from a recording of "Yellow Submarine." Highly similar frames produce white pixels, dissimilar frames produce black, and independent frames gray. From left to right, the SSMs represent (top row) MFCCs, chroma, FPs, and (bottom row) RMS, and periodicity histograms. The black lines indicate the boundaries of the structural annotation seen in Figure 1.

Each of the above features gives, for every frame, a vector of some length. We transformed the values in two ways: first, each vector dimension was standardized over the length of the piece to have zero mean and unit variance. Since no frame-wise normalization was used for any feature, this ensures the variance in each dimension is weighed equally, ensuring that repetitions in low-magnitude signals are detected albeit at the cost of some additional noise. The features were then smoothed in time; for the SSMs in Figure 2, a 10-second moving-average filter was used. Finally, the SSMs were calculated using cosine similarity.

2.2 Combining SSMs

Suppose we have an annotation for a song's structure, expressed as an SSM like in Figure 1, and want to find how best to explain it in terms of SSMs generated from the song's acoustic features, like those in Figure 2. We know that summing the feature matrices is a useful technique, and since we have the annotation we could try to calculate the optimal linear combination of feature matrices to reconstruct the annotation. This would provide a relative weight to each feature corresponding to its salience with respect to the entire song. However, knowing that the salience of different features can vary throughout the piece, we may wish to explain the annotation section by section.

Previous approaches to decomposing SSMs focused on discovering the structure of recordings, and hence used estimation techniques such as singular value decomposition [9] or NMF [16]. However, since our goal is to learn about the relationship between the known structure and the recording, we can use straightforward optimization techniques. We propose to use a quadratic program (QP), a generic formulation of an optimization problem, to find the optimal combination of feature-derived SSMs to derive the annotation-derived SSM in a piecewise fashion.

To illustrate the approach, we use a very simple example: suppose we have a piece with structure ABC, where the last section is twice as long as the previous sections (in general we may have s sections). The annotation matrix N could be:

$$N = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}$$

Recall that –1 indicates a contrasting pair and +1 indicates an identical pair. We would like to explain the reason behind each section using two features: a harmony-based feature and a timbre-based feature. (Again, in general we may have f features.) Suppose the pitch content of the song is identical for sections A and B, and the timbre is identical for sections B and C. For example, A, B and C could be the introduction, verse and chorus of a pop song, where the instrumentation changes after the introduction but stays constant thereafter, and where the pattern of chords only changes at the chorus. The two matrices F_1 and F_2, derived from the harmonic and timbral audio features, respectively, would be:

$$F_1 = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \qquad F_2 = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{bmatrix}$$

Neither matrix equals the true underlying structure N, but we would like to reconstruct the annotation matrix using subsets of the feature matrices that correspond to the annotated sections. Since there are three sections and two feature matrices, there are six available components, shown below. We can generate these by applying three "masks," one for each section, to each feature. $M_{i,j}$, the element-wise product of the j^{th} mask with F_i, will show how the j^{th} section relates to the other sections with respect to the i^{th} feature.

$$M_{1,1} = \begin{bmatrix} 1 & .5 & -.5 & -.5 \\ .5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \end{bmatrix} \qquad M_{2,1} = \begin{bmatrix} 1 & -.5 & -.5 & -.5 \\ -.5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \end{bmatrix}$$

$$M_{1,2} = \begin{bmatrix} 0 & .5 & 0 & 0 \\ .5 & 1 & -.5 & -.5 \\ 0 & -.5 & 0 & 0 \\ 0 & -.5 & 0 & 0 \end{bmatrix} \qquad M_{2,2} = \begin{bmatrix} 0 & -.5 & 0 & 0 \\ -.5 & 1 & .5 & .5 \\ 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & 0 \end{bmatrix}$$

$$M_{1,3} = \begin{bmatrix} 0 & 0 & -.5 & -.5 \\ 0 & 0 & -.5 & -.5 \\ -.5 & -.5 & 1 & 1 \\ -.5 & -.5 & 1 & 1 \end{bmatrix} \qquad M_{3,2} = \begin{bmatrix} 0 & 0 & -.5 & -.5 \\ 0 & 0 & .5 & .5 \\ -.5 & .5 & 1 & 1 \\ -.5 & .5 & 1 & 1 \end{bmatrix}$$

The $M_{1,i}$ matrices correspond to how the sections interrelate with respect to harmony, and the $M_{2,i}$ matrices show the same with respect to timbre. The masks halve all of the elements in the off-diagonal sections so that the feature matrices can be reconstructed by summing the components (i.e., $\Sigma_{j=1..s}\, M_{i,j} = F_i$). We would like to find a linear combination of the component matrices $M_{i,j}$ that will approximate the annotation N as closely as possible. That is, we want the vector of coefficients $x = \{x_{1,1}, x_{1,2}, ..., x_{f,s}\}$ that minimizes the squared distance between the annotation and the reconstruction:

$$\left(\left(\sum_{j=1}^{s} \sum_{i=1}^{f} x_{i,j} M_{i,j} \right) - N \right)^2 \tag{1}$$

This problem is solvable as a quadratic program (QP) if we imagine each component matrix $M_{i,j}$ to be a single row in a larger array $\mathbf{M}$. If each $M_{i,j}$ is an n x n matrix, then letting $k = (i-1) \cdot f + j$ we can let $\mathbf{M_k}$, the k^{th} row of $\mathbf{M}$, be the horizontal concatenation of the n rows of $M_{i,j}$:

$$\mathbf{M} = \begin{bmatrix} M_{1,1_{(1,1)}} & M_{1,1_{(2,1)}} & \cdots & M_{1,1_{(1,2)}} & \cdots & M_{1,1_{(n,n)}} \\ M_{1,2_{(1,1)}} & M_{1,2_{(2,1)}} & \cdots & M_{1,2_{(1,2)}} & \cdots & M_{1,2_{(n,n)}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ M_{f,s_{(1,1)}} & M_{f,s_{(2,1)}} & \cdots & M_{f,s_{(1,2)}} & \cdots & M_{f,s_{(n,n)}} \end{bmatrix}$$

(In our example, $\mathbf{M}$ would have 6 rows and 16 columns, since there are 16 values in each of the 6 components $M_{i,j}$.)

If we similarly reshape N into a single row vector, and treat x as a column vector, then we may rewrite expression (1) as $(x^T\mathbf{M} - N)^2$. Expanding, we obtain the following expression for c, the reconstruction cost:

$$c(x) = x^T\mathbf{M}\mathbf{M}^Tx - 2N\mathbf{M}^Tx + NN^T \tag{2}$$

Here, $\mathbf{M}\mathbf{M}^T$ is a square matrix with f rows and columns, and NN^T is a constant term which can be ignored in the QP. Our goal is to minimize $c(x)$ subject to any constraints we may place on x. We set $x \geq 0$ and interpret each coefficient $x_{i,j}$ as the relevance of feature i in explaining the similarity of section j to the entire piece. The final QP formulation is:

$$\min_x \; x^T\mathbf{M}^2x - N\mathbf{M}^Tx \tag{3}$$

$$\text{such that:} \quad x_{i,j} \geq 0, \forall\; i = 1,..,f, \; j = 1,..,s$$

This is the standard form for QPs, and is quickly solvable on commercial software. All the QPs in this article were solved using the `quadprog` function in MATLAB's optimization package. The inequality is the only constraint placed on the solution; we do not enforce the typical constraint $\Sigma\, x_{i,j} = 1$ as its interpretation is unclear and we never encountered any problems with degenerate solutions.

Solving this quadratic program for our example gives $x = \{0, 0.6875, 1.0625, 1.25, 0.3125, 0\}$. The reconstruction of the annotation using these coefficients is:

$$\mathbf{M}^T x = \begin{bmatrix} 1.25 & -.44 & -1.16 & -1.16 \\ -.44 & 1.00 & -.72 & -.72 \\ -1.16 & -.72 & 1.06 & 1.06 \\ -1.16 & -.72 & 1.06 & 1.06 \end{bmatrix}$$

($\mathbf{M}^Tx$ is actually a column vector, but here we have reshaped the result into the reconstructed matrix it represents.) The largest components are $x_{1,3}$ and $x_{2,1}$; indeed, the most explanatory components are $M_{1,3}$, which perfectly shows how section C is distinguished from A and B on the basis of its harmony, and $M_{2,1}$, which shows how A differs from the others on the basis of its timbre. The coefficient $x_{1,1}$ is 0, which properly reflects that the harmony of section A is meaningless for distinguishing it from the rest of the piece, and vice versa for $x_{3,2}$. The intermediate values of $x_{1,2}$ and $x_{2,2}$ reflect that it is relatively difficult to explain the middle section with these features. Component $M_{1,2}$ distinguishes section C at the expense of conflating A and B, while $M_{2,2}$ distinguishes A and conflates B and C. Since there is a greater cost for mischaracterizing the longer section, $x_{1,2}$ is larger than $x_{2,2}$.

The reconstruction cost $c(x)$ for the above solution is 1.125, compared to the maximum cost of 16.0 which is reached when every $x_{i,j} = 0$. If instead of this section-wise approach we had used a matrix-wise approach with two components, F_1 and F_2, we would have found the optimal coefficients 0.67 and 0.33, which gives a reconstruction cost of 5.33. Hence the section-wise approach gets over four times closer to the annotation than the matrix-wise approach. More importantly, the coefficients $x_{i,j}$ reveal when in the piece the different features are most relevant for determining its structure: in this case, harmony is an unimportant feature near the beginning of the piece, but becomes important later on, and vice versa for timbre.

The section-wise QP contains all solutions to the matrix-wise QP as a subset. The solution to the section-wise QP is thus guaranteed to be at least as good, and the reduction in reconstruction cost is no surprise. While this prevents us from quantitatively evaluating the effectiveness of the section-wise QP, we may use the matrix-wise QP as a performance ceiling and evaluate the result qualitatively.

2.3 Reconstructing an annotation SSM from audio SSMs

Using the features explained in Section 2.1.1 and the annotated information described in Section 2.1.2, we can formulate a QP using the method in Section 2.2. We demonstrate this procedure for the song "Yellow Submarine." Figure 3 illustrates how the five feature-derived SSMs and seven section masks produce 35 components. Labeling the component matrices $M_{1,1}$ through $M_{5,7}$, our goal is to find the coefficients $x = \{x_{1,1}, \ldots, x_{5,7}\}$ that minimize $c(x)$. We solve the QP and illustrate the weights x in Figure 4.

The results suggest that the first verse, A_1, is best explained by a combination of its chroma and fluctuation pattern vectors, that the first chorus, B_1, is explained almost wholly by its chroma vectors, and so forth. The prominence of FPs in the first and last sections reflects the fact that the FPs were very robust to the changes that occur partway through each section: midway through A_1, a number of nautical sound effects intrude and affect the MFCCs and chroma, and the fadeout in B_3 affects the similarity of other features. Referring to the top and bottom rows of components in Figure 3, it is clear that FPs best represent the homogeneity of the first and last sections.

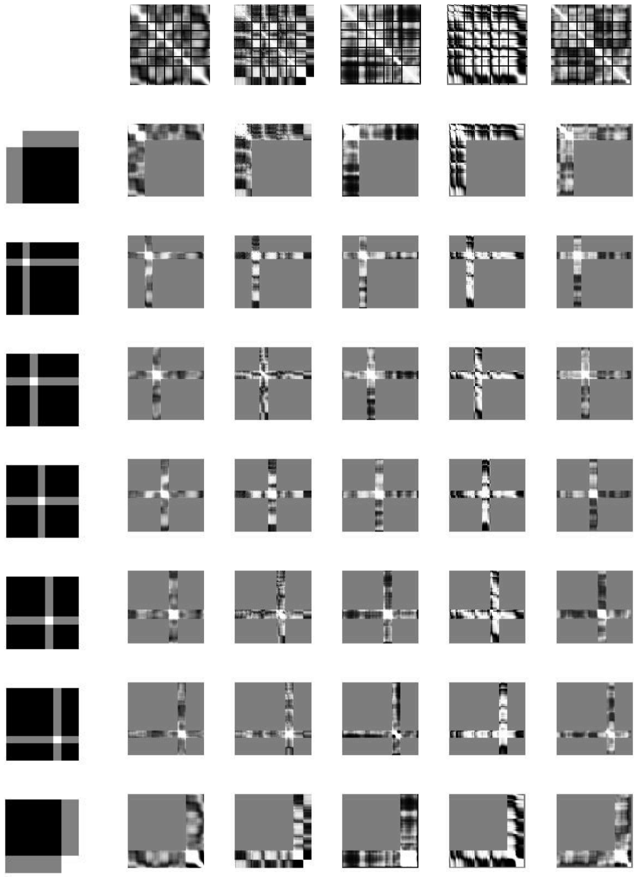

Figure 3. Illustration of matrix components. The feature matrices are given in the top row (left to right: MFCCs, chroma, FPs, RMS, and periodicity histograms); the masks in the left column represent each section in the annotation. Each component matrix is the element-wise product of a feature matrix and a mask. The feature matrices and the products are scaled from -1 (black) to +1 (white), while the masks are scaled from 0 (black) to +1 (white).

Section C_1 stands out from the piece as being best explained with a combination of timbre and tempo features. Indeed the most distinguishing feature of this section is its timbre, since it is an instrumental portion that contains many unusual sound effects like splashes and bells. Perhaps the arrhythmic nature of these sounds led the periodicity histograms to detect a strong dissimilarity with the rest of the piece.

Our method estimates connections between an analysis and the features, and our analysis of this example suggests that the connection plausibly relates to the listener's experience. However, whether the feature weights obtained by the QP actually correlate to the listener's justifications for their analysis remains a matter of conjecture. Settling this question would require paired data—annotations coupled with listener's self-reported justifications—that is not presently available, though we do plan to collect such data in the future.

2.3.1 Reconstruction cost

Subjectively, the x-values found by the QP are reasonable, but we would like to obtain some quantitative estimate of how well this method works compared to others. One measure of the quality of the output is the reconstruction cost c, which is the average squared deviation between the reconstructed matrix and the target annotation. (This is also the value of the objective function (2) at the solution found by the QP.) The maximum allowable reconstruction cost cannot exceed N^2, since this can be obtained trivially by setting $x = 0$. We can thus express the fractional reconstruction cost c/c_0, where c_0 is the cost at $x = 0$.

With this metric, we can compare the quality of different quadratic programs. To fairly estimate how much analyzing the song section by section instead of all at once improves the reconstruction, we need to run a second quadratic program: this one simply finds the coefficients $x = \{x_1, x_2, \ldots, x_f\}$ that makes the sum of the feature

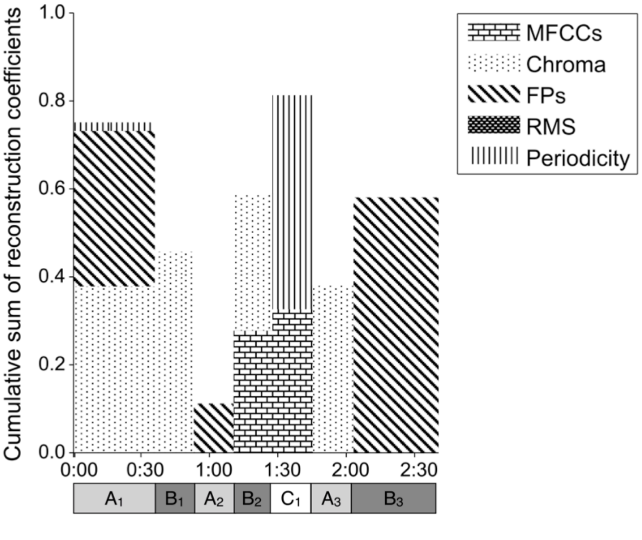

Figure 4. The optimal reconstruction coefficients for five different features for all sections of the song "Yellow Submarine." The height of each block is the value of the reconstruction coefficient for the section indicated on the *x*-axis. The annotation is given below the graph.

matrices as close to the annotation as possible. This method gives a fractional reconstruction cost of 0.81, whereas the section-wise method garnered a fractional cost of 0.68. This result is expected, since (as noted in Section 2.2) the coarse matrix-wise solution can never be better than the finer-grained solution. Still, by examining how this improvement tapers off with finer-grained formulations of the QP, as done in the next section, we can assess the limit of this method's effectiveness.

The matrices reconstructed using these two methods are pictured in Figure 5, along with a plot of the mean squared deviation from the annotation, which highlights those regions that are poorly reconstructed. The latter plots make evident that both approaches have trouble reconstruction the edges of the SSM, which describe how the very beginning and end of the song relate to the rest. The matrix-wise approach also has particular trouble reconstructing the third verse (section A_3)

2.3.2 Reconstruction using smaller sections

The section-wise approach was mathematically guaranteed to result in at least as good an approximate of the annotation as a linear combination of full matrices. We may expect even better approximations if we divide the matrix into smaller sections. However, if further segmentation is structurally irrelevant, the reductions in reconstruction cost will taper off. We repeated the previous QP using the finer segmentation of the small-scale sections as well as a "finest-scale" segmentation with segments every 2.5 seconds—shorter than the longest feature windows.

Looking at finer time scales reveals new insights: for example, it is very noticeable that section c_1 is poorly explained by all the features (Figure 6). Indeed, this small section contains a novel tune played on a brass instrument and sounds nothing like the rest of the piece. It doesn't even sound like the later sections c_2 and c_3, although their relatedness could be argued by their both containing sound effects and by their being some kind of variation of the usual A section. Also, whereas in the previous analysis (Figure 4) we saw that section C_1 was explained both by its distinct timbre and potentially confusing tempo, we can see now that each half of the section is better explained by one of these features. The first half, c_2, has less percussion than the second half and is arguably the more confusing.

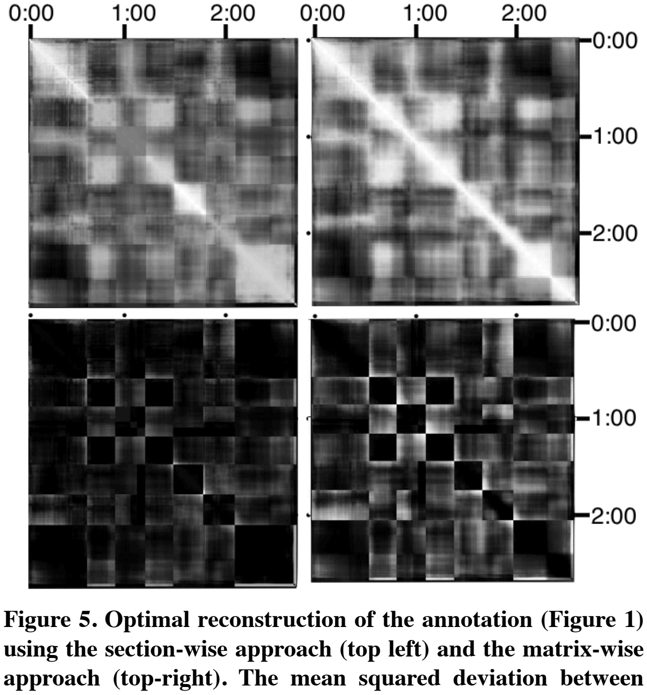

Figure 5. Optimal reconstruction of the annotation (Figure 1) using the section-wise approach (top left) and the matrix-wise approach (top-right). The mean squared deviation between each reconstructed matrix and the annotation is shown below the reconstruction. (Black pixels indicate where the reconstruction is most accurate.)

Drilling down further to 2.5-second segments (Figure 7), the result is more detailed but not necessarily more informative: the sum of the coefficients leaves a similar "skyline", indicating that approximately the same amount of information is explained in each. However, a few parts are better explained at this smaller scale: the best explanation for *a_5* and *a'* switches from FPs to

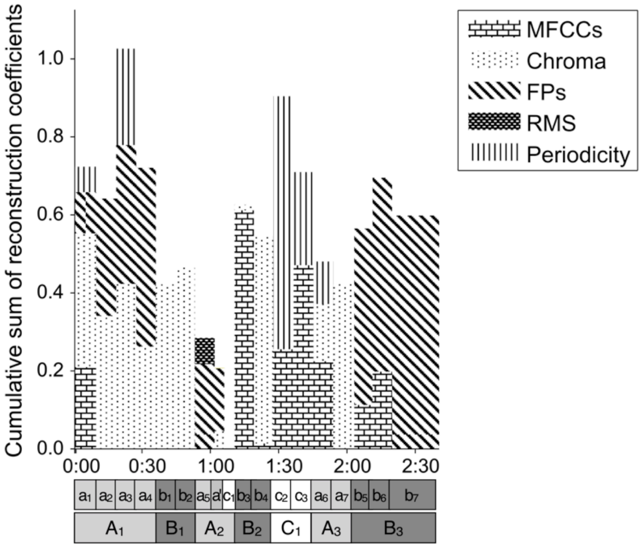

Figure 6: The optimal reconstruction coefficients for each small-scale section.

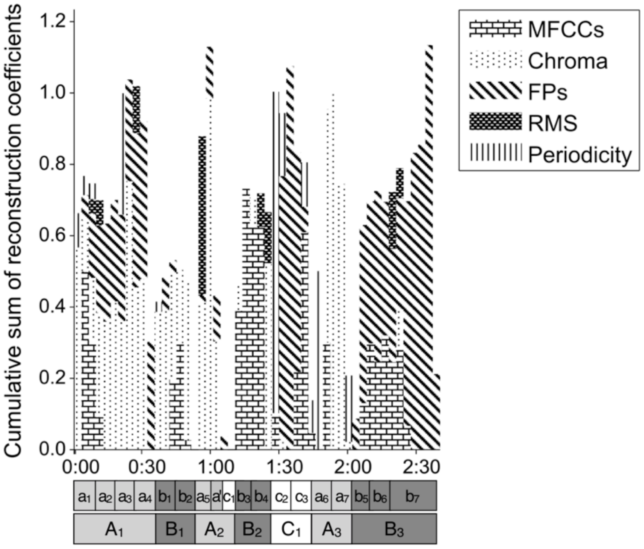

Figure 7: The optimal reconstruction coefficients for each 2.5-second frame in the piece.

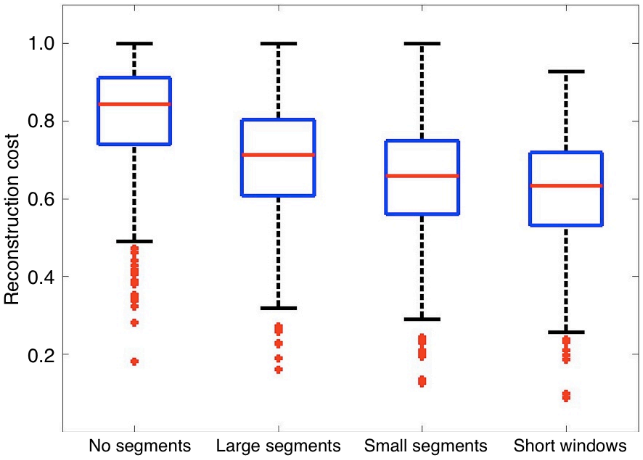

Figure 8. Boxplot of QP reconstruction costs obtained for QPs using no segmentation (matrix-wise sum) and the piece-wise decomposition at three time scales. Data include 704 recordings from the SALAMI corpus.

chroma, bringing A_2 in line with the other A sections. Also, the coefficients in b_7 are higher at this scale.

As stated earlier, finer-grained segmentations are guaranteed to lead to better reconstructions. In the above example, the fractional cost for the small-scale segmentation was 0.67 (compared to 0.68 for the large-scale segmentation), and for the short-window analysis it was 0.61. An analysis of reconstruction costs over a larger corpus reveals that this example is typical. The QP algorithm was executed on annotations for 704 recordings in the SALAMI dataset at the four levels of granularity: matrix-wise, large-scale, small-scale, and finest-scale. The reconstruction costs were computed for each and are plotted in Figure 8. We see that while large reductions in reconstruction cost are typical when moving from matrix-wise to large-scale QP formulations, there is less improvement moving from small-scale to short window, indicating diminishing returns when the segmentation proceeds beyond what the annotation contains.

3. VISUALIZING STRUCTURAL DIFFERENCES

The previous examples show that the relationship between the structure of a piece of music and its feature-derived SSMs can vary over time: repetitions in a feature that are irrelevant at one point in a piece may be foregrounded at another. Just as [15] and [25] argued that dynamic feature weighting could improve structural analysis, our examples show that dynamic feature interpretation could aid in applications based on structural information. Here, we focus on its use as a visualization tool.

For example, the data obtained by our approach may provide interesting visualizations for projects like SALAMI, which plans to execute several algorithms to annotate the musical structure for a large library [31]. To facilitate browsing in it they have developed a system to visualize each structural description with a diagram, like the annotation in the lower part of Figure 7 [5]. Providing the section-wise estimates of which features are estimated to be most salient could enrich these diagrams.

Our method of decomposing annotations is also suited to comparing annotations prepared by different listeners. We illustrate this with two examples. The first shows how a single large difference between two annotations is reflected in the reconstruction coefficients, and the second demonstrates the power of the approach to reconstruct greatly divergent analyses from the same set of components.

3.1 Investigating a single difference

As noted in the introduction, it is common for two listeners to analyze the same piece of music differently, and this may be because they are paying attention to different acoustic features. Our analysis method allows one to investigate the differences between two analyses, showing how each may have arisen by emphasizing certain features over others at certain times.

We illustrate this potential by reconstructing two different annotations of the piece “Garrotin”, a solo flamenco guitar piece recorded by Chago Rodrigo. Two SALAMI annotators gave analyses that were similar overall (Figure 9): the piece begins and ends with many repetitions of the same main melodic gesture, and the middle of the song (roughly 0:40 to 1:00) consists of a number of different melodic episodes separated by short calls back to the main theme.

The analyses differ mainly in their treatment of the middle section: the first listener interprets it as a single episode, while the second listener analyzes it as two distinct episodes. The feature SSMs (Figure 10) show that this portion of the piece is quite self-similar with respect to pitch and timbre, but an internal contrast to the section is revealed in the FP-derived SSM at the shorter time scale (5 seconds). This difference is reflected in the solutions to the QP (Figure 11): the middle section of the piece is reconstructed best by chroma features when kept as one section, but reconstructed better by FPs when divided in two.

3.2 Reconstructing dissimilar analyses

In the previous example, a small disagreement was investigated. What if listeners have vastly different interpretations—is it still possible to find QP solutions that justify each interpretation equally well? We look at such an example now: “As the Bell Rings the Maypole Spins”, by the World music band Dead Can Dance. In the song, a singer and bagpipes play a series of reels, and the pattern of reels repeats a few times before a long repetitive coda section ends the piece. The stark difference between the two annotations is apparent from the SSMs (Figure 12, top row). The first listener has identified a sequence of three reels as a self-contained repeating group, leading to large off-diagonal blocks in the middle of the SSM. The second listener has not identified these larger groupings, but does indicate that many of the reels are identical or similar to the coda section (from 3:40 onward), resulting in a series of thin bars in the SSM.

Despite the differences in the annotations, a QP using five features and three time scales has reconstructed the annotations qualitatively well (Figure 12, middle row). The fractional reconstruction costs for the first and second annotations, when using the large-scale segmentation, are .63 and .57, compared to .74 and .76 when using no segmentation.

Examining the reconstruction coefficients (Figure 13), we can observe that the two solutions depend on fairly distinct sets of

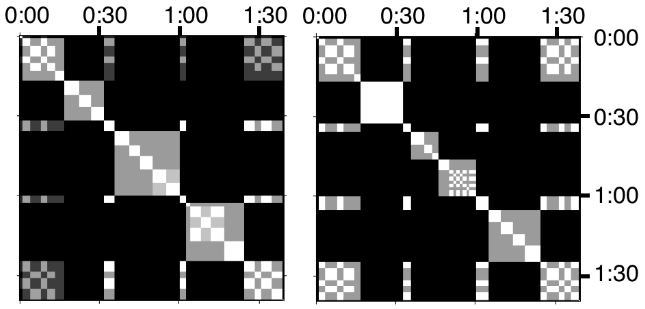

Figure 9. Annotations by two listeners for "Garrotin" (salami_id: 842).

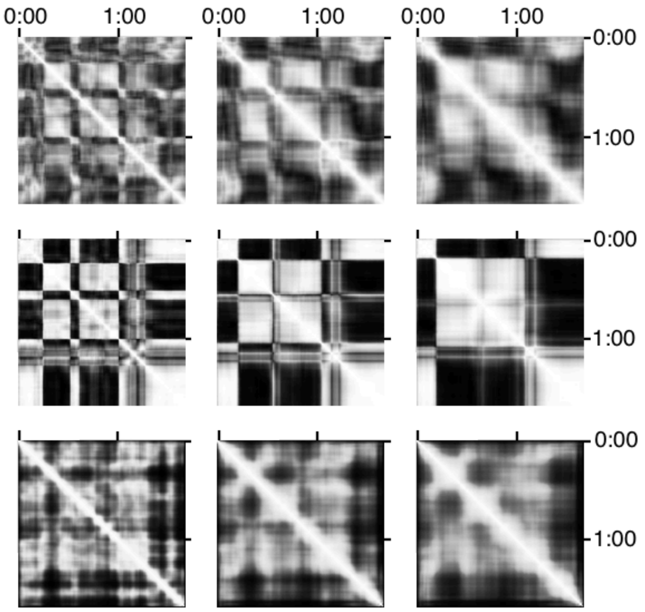

Figure 10. Feature matrices for the reconstruction of "Garrotin." In this example, three features are used: MFCCs (top row), chroma (middle row), and FPs (bottom row), as well as three smoothing window sizes: 5, 10 and 15 seconds (left, middle and right columns, respectively).

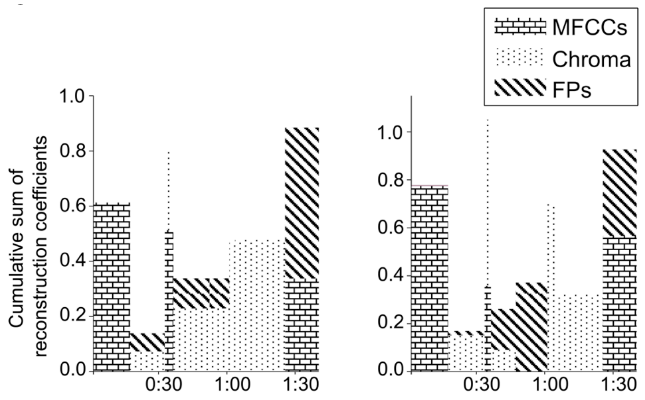

Figure 11. The optimal reconstruction coefficients for three different features for large sections of the song "Garrotin." The *x*-axis represents time. The coefficients shown here are the average of the coefficients for the three different time scales.

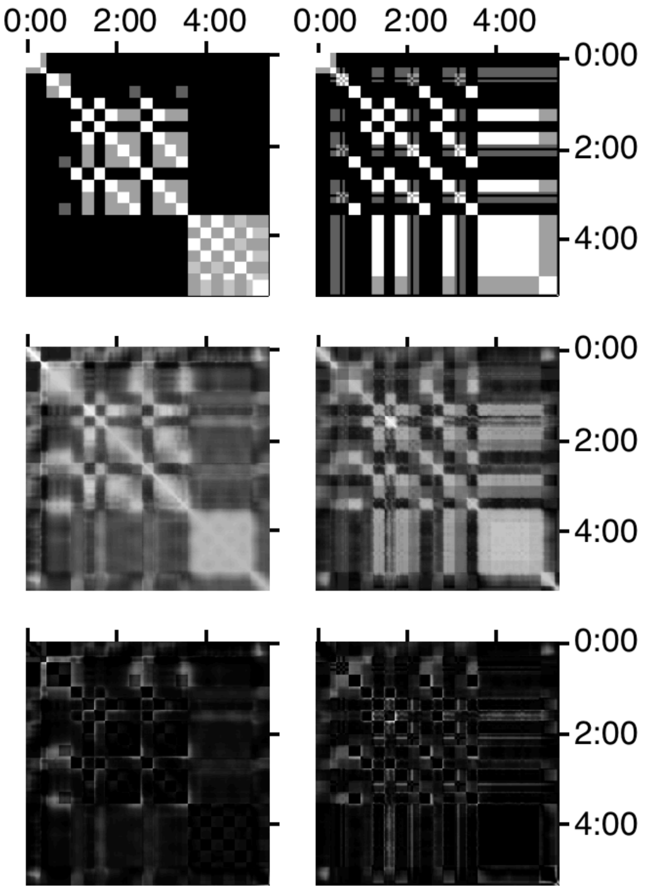

Figure 12. Top row: SSMs from two annotations for "As the Bell Rings the Maypole Spins", by Dead Can Dance (salami_id: 860). Middle row: reconstructed matrices. Bottom row: mean squared reconstruction error.

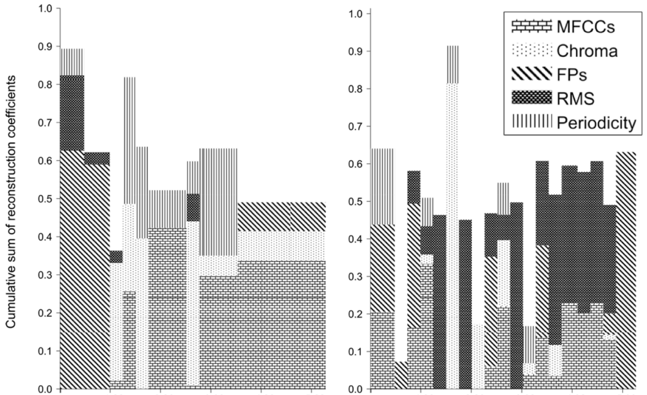

Figure 13. The optimal reconstruction coefficients for five features for large sections of the song "As the Bell Rings the Maypole Spins." The coefficients shown here are the average of the coefficients for the three different time scales.

features. The first annotation is explained best by FPs in the first section (up to 1:00), and thereafter mainly by MFCCs. By contrast, much of the second annotation is explained best by RMS. Both solutions involve a mixture of features between 1:00 and 2:30, but the mixtures are distinct.

4. CONCLUSION & FUTURE WORK

We have introduced the problem of estimating the relevance of different acoustic features to different sections of a structural analysis, and proposed a method of solving the problem based on quadratic programming. The approach is founded on the intuition that while acoustically-derived SSMs may not always reflect the true structure of a piece, components of SSMs for specific features may explain parts of the true structure. The data explored here relate to listeners' analyses, and it is possible that our method reveals insights into how listeners analyzed the pieces. However, to establish a correlation between the SSMs and the listeners' procedure for analyzing the piece would require new experimental data, the collection of which remains future work

The method presented is a general one and the parameters of the QP, such as the mask shape, could be modified in interesting ways to suit different problems. For example, if the masks (as in Figure 4) were altered to only include the upper left portion of the SSM, a future-agnostic analysis would result. The annotations considered here were all produced after the music had been fully heard, but response data that reflected one's real-time perception of structure would perhaps best be analyzed with such a future-agnostic framework.

Similarly, a QP using masks that emphasized the main diagonal might model a listening experience with less long-term memory. Such an analysis would provide a solution that explained local similarities and contrasts at the expense of more distant sections. This version would be useful if, for instance, once considered the acoustic similarity between the very beginning and end of a piece to be unimportant. (The more narrowly one focuses on the diagonal axis, the more this method resembles the novelty function calculation proposed by [8]. Except, instead of correlating the diagonal axis with a checkerboard kernel, we would be correlating it with the ground truth annotation.)

One important caveat with our approach is that it is crucial that the annotation used properly reflect the information that one seeks to explain. In this article we have taken for granted the "states" hypothesis of the annotations, to use the terminology of [28]. That is, we have assumed that a section with a single label is homogenous, and uniformly distinct from any differently-labeled section. It would be useful to extend our work to account for "sequence" interpretations of annotations, in which a section B is presumed to be a heterogeneous sequence of events that recurs exactly whenever B repeats. The use of structural information at multiple timescales, described in Section 2.1, is intended to mitigate this shortcoming, since in practice the short-time scale annotation often charts a "sequence"-like path through the blocks of the large-scale annotation. Implementing a sequential similarity metric is not straightforward since two annotated sections can have the same label but very different lengths, and an automatic alignment algorithm such as [20] would have to be incorporated..

5. ACKNOWLEDGMENTS

This research was supported in part by the Social Sciences and Humanities Research Council of Canada and by a QMUL EPSRC Doctoral Training Account studentship. We would like to thank the anonymous reviewers for their generous help in improving this submission.

6. REFERENCES

[1] Bartsch, M. and Wakefield, G. 2001. To catch a chorus: using chroma-based representations for audio thumbnailing. *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, USA). 15–18.

[2] Bruderer, M., McKinney, M., and Kohlrausch, A. 2009. The perception of structural boundaries in melody lines of Western popular music. *Musicae Scientae*. 13, 2, 273–313.

[3] Clarke, E. F. and Krumhansl, C. L. 1990. Perceiving musical time. *Music Perception*. 7, 3, 213–251.

[4] Eckmann, J. P., Kamphorst, S. O., and Ruelle, D. 1987. Recurrence plots of dynamical systems. *Europhysics Letters*, 5, 9, 973–977.

[5] Ehmann, A. F., Bay, M., Downie, J. S., Fujinaga, I., and De Roure, D. 2011. Exploiting music structures for digital libraries. In *Proceeding of the International ACM/IEEE Joint Conference on Digital Libraries* (Ottawa, Canada). 479–480.

[6] Eronen, A. 2007. Chorus detection with combined use of MFCC and chroma features and image processing filters. In *Proceedings of the International Conference on Digital Audio Effects* (Bordeaux, France). 229–236.

[7] Foote, J. 1999. Visualizing music and audio using self-similarity. In *Proceedings of the ACM International Conference on Multimedia* (New York, NY, USA). 77–80.

[8] Foote, J. 2000. Automatic Audio Segmentation using a Measure of Audio Novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 452–455.

[9] Foote, J. and Cooper, M. 2003. Media segmentation using self-similarity decomposition. In *Proceedings of the SPIE: Storage and Retrieval for Media Databases* (Santa Clara, CA, USA). 167–175.

[10] Frankland, B. and Cohen, A. 2004. Parsing of melody: Quantification and testing of the local grouping rules of Lerdahl and Jackendoff's *A Generative Theory of Tonal Music*. *Music Perception*. 21, 4, 499–543.

[11] Goto, M. 2006. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech and Language Processing*. 14, 5, 1783–1794.

[12] Grosche, P., Serrà, J., Müller, M., and Arcos, J. L. 2012. Structure-based audio fingerprinting for music retrieval. In *Proceedings of the International Conference on Music Information Retrieval* (Porto, Portugal). 55–60.

[13] Hargreaves, S., Klapuri, A., and Sandler, M. 2012. Structural segmentation of multitrack audio. *IEEE Transactions on Audio, Speech, and Language Processing*. 20, 10, 2637–2647.

[14] Jehan, T. 2005. Hierarchical multi-class self similarities. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, USA). 311—314.

[15] Kaiser, F. and Peeters, G. 2012. Adaptive temporal modeling of audio features in the context of music structure segmentation. In *International Workshop on Adaptive Multimedia Retrieval* (Copenhagen, Denmark).

[16] Kaiser, F. and Sikora, T. 2010. Music structure discovery in popular music using non-negative matrix factorization. In *Proceedings of the International Society for Music Information Retrieval Conference* (Utrecht, The Netherlands). 429–434.

[17] Landone, C. Gasser, M., Cannam, C., Harte, C., Davies, M., Noland, K., Wilmering, T., Xue, W., and Zhou, R. 2011. QM Vamp Plugins. Available: <http://isophonics.net/QMVampPlugins>, accessed 1 October 2012.

[18] Marolt, M. 2006. A mid-level melody-based representation for calculating audio similarity. In *Proceedings of the International Conference on Music Information Retrieval* (Victoria, Canada). 280–285.

[19] Mauch, M., Noland, K., and Dixon, S. 2009. Using musical structure to enhance automatic chord transcription. In *Proceedings of the International Society for Music Information Retrieval Conference* (Kobe, Japan). 231–236.

[20] Müller, M. & Appelt, D. 2008. Path-constrained partial music synchronization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Las Vegas, NV, USA). 65–68.

[21] Ong, B. 2007. Structural Analysis and Segmentation of Music Signals. Ph.D. dissertation, University Pompeu Fabra.

[22] Pampalk, E. 2004. A Matlab toolbox to compute similarity from audio. In *Proceedings of the International Conference on Music Information Retrieval* (Barcelona, Spain). 254–257.

[23] Pampalk, E., Dixon, S., and Widmer, G. 2004. Exploring music collections by browsing different views. *Computer Music Journal*, 28, 2, 49–62.

[24] Pampalk, E., Rauber, A., and Merkl, D. 2002. Content-based organization and visualization of music archives. In *Proceedings of the ACM International Conference on Multimedia* (Juan les Pins, France). 570–579.

[25] Parry, R., and Essa, I. 2004. Feature weighting for segmentation. In Proceedings of the International Conference for Music Information Retrieval (Barcelona, Spain).

[26] Paulus, J. & Klapuri, A. 2006. Music structure analysis by finding repeated parts. In *Proceedings of the ACM Workshop on Audio and Music Computing Multimedia* (New York, NY, USA). 59–68.

[27] Paulus, J., Müller, M., and Klapuri, A. 2010. Audio-based music structure analysis. In *Proceedings of the International Society for Music Information Retrieval Conference* (Utrecht, The Netherlands). 625–636.

[28] Peeters, G. 2004. Deriving musical structures from signal analysis for music audio summary generation: "Sequence" and "State" approach. In *Computer Music Modeling and Retrieval 2771*. Springer Berlin / Heidelberg, 169–185.

[29] Peeters, G. 2007. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proceedings of the International Conference on Music Information Retrieval* (Vienna, Austria). 35–40.

[30] Shiu, Y., Jeong, H., and Kuo, C.-C. J. 2006. Similarity matrix processing for music structure analysis. In *Proceedings of the ACM Workshop on Audio and Music Computing Multimedia* (New York, NY, USA). 69–76.

[31] Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D., and Downie, J. S. 2011. Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference* (Miami, FL, USA). 555–560.