Unsupervised Discovery of Temporal Structure in Music

Ron J. Weiss, *Member, IEEE*, and Juan Pablo Bello, *Member, IEEE*

***Abstract*—We describe a data-driven algorithm for automatically identifying repeated patterns in music which analyzes a feature matrix using shift-invariant probabilistic latent component analysis. We utilize sparsity constraints to automatically identify the number of patterns and their lengths, parameters that would normally need to be fixed in advance, as well as to control the structure of the decomposition. The proposed analysis is applied to beat-synchronous chromagrams in order to concurrently extract recurrent harmonic motifs and their locations within a song. We demonstrate how the analysis can be used to accurately identify riffs in popular music and explore the relationship between the derived parameters and a song's underlying metrical structure. Finally, we show how this analysis can be used for long-term music structure segmentation, resulting in an algorithm that is competitive with other state-of-the-art segmentation algorithms based on hidden Markov models and self similarity matrices.**

***Index Terms*—Convolutive non-negative matrix factorization (NMF), music structure analysis, sparse priors.**

I. Introduction

Repetition is widely acknowledged to play a fundamental role in music, with many common musical terms, such as riff, groove, motive, tempo, meter, or section, largely defined as a function of the presence or absence of recurrent patterns. In its many guises, repetition has been linked to the coherence and intelligibility of musical works, and features prominently in the most influential theories of music analysis, often associated with notions of structural organization and form [1]–[3].

While particularly strong in popular music, this prevalence of repetition is ubiquitous across periods, styles and traditions. This is exemplified by the recurrent riffs and sections of both punk and salsa music; the recapitulation of themes and motives, often in different keys and tempos, which are as common in bebop as in music from the classical and romantic periods; and the preference for repetitive rhythmic patterns manifest in western African music and electronica. It is therefore clear that the characterization of repetitive patterns and their temporal organization is central to the analysis and understanding of most music.

Several approaches have been proposed for the discovery of repetitive patterns in symbolic representations of music. Examples include the use of string matching techniques and models of common listening strategies on isolated melodic lines [4]–[6]; and of multiple viewpoints and the geometrical analysis of multi-dimensional representations for polyphonic music [7], [8]. However, discovering repetitive patterns from audio signals poses a significantly more challenging problem, as they have to be untangled from the noisy mix resulting from the interaction between musicians, instruments, and the recording process. In this context, the automatic extraction of even the most basic information in a score, such as the start time, pitch, and duration of notes, has proven a difficult task for which a canonical solution has yet to be found.

In music signal processing research, work on the characterization of repetitive patterns is usually framed in the context of music structure analysis (for a detailed review see [9]). Example strategies include the use of agglomerative clustering [10]; hidden Markov models (HMMs) combined with simple aggregation [11], string matching [12], k-means clustering [13], and Bayesian clustering of state histograms [14]. The most popular approach, however, is based on the analysis of self-similarity matrices [15], where repetitions are characterized by diagonals or blocks of small distance values. This property has been exploited for tasks as diverse as visualization, rhythmic analysis, automatic summarization and thumbnailing, chorus detection, annotation, synchronization, and long-term segmentation [16]–[20]. With a few exceptions [17], [21], [22], the emphasis of this research has been on locating repetitions rather than on extracting of characteristic, repetitive patterns. The utility of extracting such patterns is illustrated by previous research on detecting motif occurrences across a collection [23] and cover-song retrieval based on feature sub-sequences [24].

In this paper, we describe a novel approach for the automatic extraction and localization of repeated patterns in music audio. The approach is based on sparse shift-invariant probabilistic latent component analysis [25] (SI-PLCA), a probabilistic variant of convolutive non-negative matrix factorization (NMF) [26]. The algorithm treats a musical recording as a concatenation of a small set of short, repeated patterns, and is able to simultaneously estimate both the patterns and their repetitions throughout the song. We show how sparse prior distributions can be used to learn the number of patterns and their respective lengths, minimizing the number of parameters that must be specified exactly in advance.

Manuscript received September 29, 2010; revised February 20, 2011; accepted April 01, 2011. Date of publication April 21, 2011; date of current version September 16, 2011. This work was supported in part by the National Science Foundation under Grant IIS-0844654 and in part by the Institute of Museum and Library Services under Grant LG-06-08-0073-08. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Meinard Müller.

R. J. Weiss was with the Music and Audio Research Laboratory (MARL), New York University, New York, NY 10012 USA. He is now with Google, Inc., New York, NY 10011 USA (e-mail: ronw@ee.columbia.edu).

J. P. Bello is with the Music and Audio Research Laboratory (MARL), New York University, New York, NY 10012 USA (e-mail: jpbello@nyu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2011.2145356

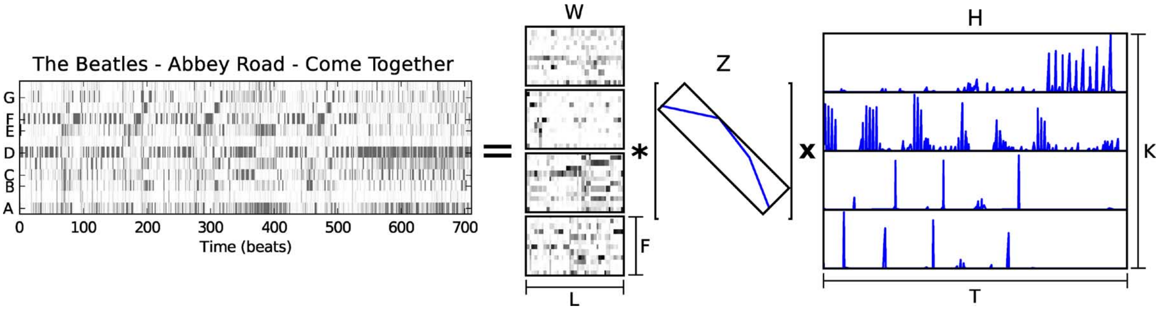


Fig. 1. Demonstration of the SI-PLCA analysis of a beat-synchronous chromagram. The decomposition was initialized with $L = 40$, and $K = 10$ with $\alpha_z = 0.98$, and no sparsity on W_k or $\mathbf{h}_k^T$. The parameter estimation algorithm pruned out most of the initial bases due to the sparse prior on $\mathbf{z}$, converging on only four bases.

Here we extend the capabilities of the baseline model, first described in [27], to be able to identify instances of a harmonic pattern in the presence of complex variations, such as key modulations. This is accomplished by extending the model to support two dimensional shift-invariance, a technique that has previously been used for musical source separation of log-frequency spectral features [28]. We explore the proposed algorithm's ability to identify repeated motifs present within a song, and demonstrate that it accurately captures rhythmic structure, i.e., a song's time signature, as well. Finally, we explore the application of this approach to long-term segmentation of musical pieces.

The remainder of this paper is organized as follows. Section II reviews the proposed analysis based on SI-PLCA and describes its relationship to NMF. Sections III and IV describe prior distributions over the SI-PLCA parameters and the expectation maximization algorithm for parameter estimation. Sections Vto VII discuss how the analysis can be used to extract the repetitive structure of music on different scales. Section V discusses how SI-PLCA can be used for motif finding, Section VI explores the relationship between the analysis and musical meter, and Section VII discusses how the proposed analysis can be used for structure segmentation. Finally, we conclude with a discussion of the limitations of the proposed approach and future work in Section VII.

II. Signal Model

A. Feature representation

For our analysis we use chroma features to represent the harmonic content of music audio. These features summarize the signal energy present in each of the 12-pitch classes of the chromatic scale. We use the implementation in [29], which averages features within automatically detected beat segments. Furthermore, the features for each beat are normalized such that the maximum energy is one. The process is designed to minimize the influence of timbre, tempo, and dynamic variations on the rest of our analysis.

The resulting representation consists of an $F = 12$ pitch classes by T beats feature matrix V. An example is shown in Fig. 1. Analysis of these beat-synchronous chroma features identifies repeated motifs in the form of chord patterns.

B. From NMF to PLCA

Conventional non-negative matrix factorization (NMF) [30] decomposes a non-negative matrix V into the product of two non-negative matrices W and H:

$$V \approx WH \tag{1}$$

where the columns of W represent basis vectors used repeatedly throughout V and the rows of H represent the activations of each basis. In the context of audio analysis, if V represents a time–frequency decomposition of an audio signal, each column of W can be thought of as a frequency template used repeatedly throughout V, and each row of H can be thought of as the activations of the corresponding basis in time. Although the focus of this paper is the analysis of chroma features, the method is equally applicable to any non-negative time-frequency representation such as a magnitude spectrogram.

Probabilistic Latent Component Analysis (PLCA) [25], [31] recasts NMF in a probabilistic framework, reminiscent of the Probabilistic Latent Semantic Analysis algorithm [32] used for text topic modeling. PLCA represents each column of W and each row of H as multinomial probability distributions and adds an additional distribution over the set of bases, i.e., a mixing weight. The decomposition can be rewritten in NMF terms as follows:

$$V \approx WZH = \sum_{k=0}^{K-1} z_k \mathbf{w}_k \mathbf{h}_k^T \tag{2}$$

where $Z = \operatorname{diag}(\mathbf{z})$ is a diagonal matrix of mixing weights $\mathbf{z}$ and K is the rank of the decomposition, i.e., the number of bases in W. Contrary to standard NMF, each of V, $\mathbf{w}_k$, $\mathbf{z}$, and $\mathbf{h}_k^T$ are normalized to sum to 1 since they correspond to probability distributions. Therefore, the decomposition can also be written as a factorization of the distribution as follows:

$$V = P(f,t) \approx \sum_k P(k)P(f|k)P(t|k) \tag{3}$$

where $P(k) = z_k$, $P(f|k) = w_{kf}$, $P(t|k) = h_{kt}$, and $f \in [0, F)$, $t \in [0, T)$ index into the rows and columns of V, respectively.

The normalization of the parameters to form distributions removes the scale indeterminacy between W and H present in conventional NMF. The probabilistic foundation furthermore makes for a convenient framework for imposing constraints on the parameters $\mathbf{w}_k$, $\mathbf{h}_k^T$, and $\mathbf{z}$ through the use of prior distributions. This will be discussed in detail in Section III.

C. Adding Shift-Invariance

In [25], Smaragdis, *et al.* describe a shift-invariant extension to the PLCA model which allows for *convolutive* bases. Unlike the single beat bases $\mathbf{w}_k$ described in Section II-B, each SI-PLCA basis is expanded to form a fixed duration template W_k containing L beats. Therefore, the $F \times K$ matrix W becomes an $F \times L \times K$ tensor $\mathcal{W}$, and the normalized basis $\mathbf{w}_k$ becomes a normalized matrix W_k. The factors $\mathcal{W}$ and H are combined via a convolution operation instead of matrix multiplication in a process analogous to the right side of (2):

$$V \approx \sum_k z_k W_k * \mathbf{h}_k = \sum_{k,\tau} z_k \mathbf{w}_{k\tau} \overset{\rightarrow\tau}{\mathbf{h}_k^T} \tag{4}$$

where $\overset{\rightarrow t}{\mathbf{x}}$ shifts $\mathbf{x}$ t places to the right and $\tau \in [0, L)$ indexes into the columns of W_k. Mirroring (3), the probabilistic interpretation of (4) can be written as follows:

$$P(f,t) \approx \sum_{k,\tau} P(k)P(f,\tau|k)P(t-\tau|k). \tag{5}$$

Fig. 1 shows an example SI-PLCA decomposition of a chromagram using $K = 4$ basis patterns of length $L = 40$ beats.

D. Two Dimensional Shift Invariance

A useful property of the chroma representation is that transpositions into different keys correspond to vertical rotations of the corresponding features in V. This motivates the addition of vertical shift invariance into the model described in the previous section. We accomplish this by expanding the activations H into a $K \times R \times T$ tensor $\mathcal{H}$, and expanding the per-basis activations $\mathbf{h}_k^T$ to form a matrix H_k whose rows correspond to different vertical rotations, i.e., relative key transpositions, of the corresponding basis W_k. R corresponds to the maximum allowed rotation of W_k, and is typically equal to F to allow the model to detect all possible key transpositions.

This decomposition can be written as follows:

$$V \approx \hat{V} = \sum_{k,\tau,r} z_k \overset{r}{\uparrow} \mathbf{w}_{k\tau} \overset{\rightarrow\tau}{\mathbf{h}_{kr}^T} \tag{6}$$

where $\overset{r}{\uparrow}\mathbf{x}$ *circularly* shifts $\mathbf{x}$ r places upward and $\mathbf{h}_{kr}^T$ corresponds to the rth row of H_k. Similar to (5), the probabilistic interpretation of (6) can be written as follows:

$$P(f,t) \approx \sum_{k,\tau,r} P(k)P(f-r,\tau|k)P(t-\tau,r|k) \tag{7}$$

Note that the models described in Sections II-B and II-C are special cases of that described here, using $L = R = 1$ and $R = 1$, respectively. We will therefore utilize the notation from this section throughout the remainder of this paper, even in situations where $R = 1$. The parameters W_k, $\mathbf{z}$, and H_k are estimated from the feature matrix V iteratively using an expectation maximization algorithm. This will be discussed in detail in Section IV.

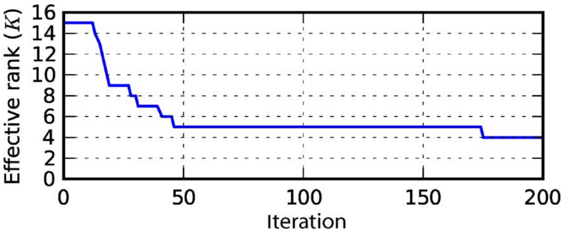


Fig. 2. Typical behavior of the automatic relevance determination process using a sparse prior on $\mathbf{z}$. The initial rank of the decomposition is set to $K = 15$, and as the estimation algorithm iterates it is pruned down to a final effective rank (the number of bases with nonzero z_k) of 4.

III. Sparse Prior Distributions

A common strategy used throughout the NMF literature to learning parsimonious, parts-based decompositions is to favor sparse settings for $\mathcal{W}$ and $\mathcal{H}$, i.e., settings containing many zeros [33]. Sparse solutions can be encouraged when estimating the parameters in (6) by imposing constraints using an appropriate prior distribution. In the following sections we describe how sparsity can be used to automatically learn the number and length of the repeated motifs within a song, and to favor solutions composed of bases that are easy to interpret.

A. Learning the Number of Patterns K

The Dirichlet distribution is conjugate to the multinomial distributions W_k, $\mathbf{z}$, and H_k, making it a natural choice for a prior. The Dirichlet prior on $\mathbf{z}$ has the following form:

$$P(\mathbf{z}|\alpha_z) \propto \prod_k z_k^{\alpha_z - 1}, \quad \alpha_z \geq 0 \tag{8}$$

where the hyperparameter α_z is fixed across all K components. If $\alpha_z < 1$ this prior favors solutions where the distribution is sparse.

By forcing $\mathbf{z}$ to be sparse, the learning algorithm attempts to use as few bases as possible. This enables an automatic relevance determination strategy in which 1) the algorithm is initialized to use many bases, i.e., K is set to a large value, and 2) the sparse prior on $\mathbf{z}$ prunes out bases that do not contribute significantly to the reconstruction of V. Only the most relevant patterns "survive" to the end of the parameter estimation process, as is shown in the example in Fig. 2. This approach is useful because it removes the need to specify the exact rank of the decomposition K in advance. The parameter estimation simply learns the underlying number of patterns needed to accurately reconstruct the data. A similar approach to automatically determining the rank of a standard NMF decomposition is described in [34].

B. Learning the Pattern Length L

Different patterns within the same piece often have different intrinsic lengths, e.g., if a song's chorus is based on a shorter riff than the verse or if its time signature changes. Therefore, it is useful to automatically identify the length of each basis independently instead of using a fixed length across all bases.

We employ a similar strategy to that described in Section III-A by setting L to an upper bound on the expected pattern length and constructing a structured prior distribution

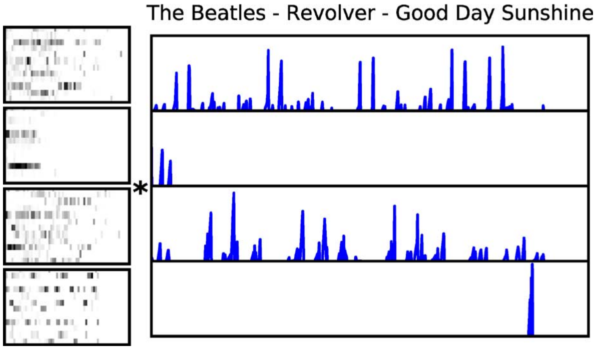


Fig. 3. Demonstration of the SI-PLCA decomposition of a chromagram using $L = 60$ and sparsity in all parameters ($\alpha_z = 0.98$, $c = 16$, $m = -10^{-7}$, and $\beta_h = 10^{-2}$).

that encourages the use of shorter bases. This is accomplished using a Dirichlet prior across the rows of W_k with a parameter that depends on the time position τ within each basis:

$$P(W_k|\alpha_{w\tau}) \propto \prod_\tau \prod_f w_{kf\tau}^{\alpha_{w\tau}-1} \tag{9}$$

$\alpha_{w\tau}$ is constructed as a piecewise function which is uninformative, i.e., imposes no constraints, for small τ and then becomes increasingly sparse:

$$\alpha_{w\tau} = \begin{cases} 1, & \tau < c \\ 1 + m(\tau - c), & \tau \geq c \end{cases} \tag{10}$$

where $c \in [1, L)$ is the beat at which the prior becomes active, i.e., the minimum pattern length, and $0 \leq m \ll 1$ is the sparseness penalty. This prior only affects patterns longer than c beats with a penalty that increases with the pattern length.

An example of the effect of this prior is shown in Fig. 3. Most of the information in the second basis is contained within the first 16 columns, while the other bases have effective lengths between 30 and 40 beats.

C. Sparse Activations

It is often worthwhile to apply sparsity constraints on H_k to obtain more informative patterns and to avoid converging on sub-optimal parameter settings. The rationale is that if most of the activations in H_k are zero, then more of the information in V will be captured by W_k since bases would be less likely to overlap during reconstruction.

In general, sparse activations promote the identification of more informative patterns in W_k at the cost of reduced time resolution in H_k. This is illustrated by the example in Fig. 1. The second basis pattern is relatively sparse, while the corresponding element of $\mathcal{H}$ contains many nonzero entries. In fact, the spacing between adjacent activations in $\mathbf{h}_1$ is smaller than the length of the pattern; i.e., it is continually mixed with delayed versions of itself. The pattern repeats about every eight beats, roughly corresponding to the underlying meter. In contrast, the bottom two bases are significantly more dense while the corresponding elements of $\mathcal{H}$ contain only about four peaks. The sparsity parameters over $\mathcal{H}$, in combination with those of $\mathcal{W}$ control the trade-off between these qualitatively different solutions. A sparse $\mathcal{H}$ leads to more musically meaningful bases that are exactly repeated throughout the piece, while a sparse $\mathcal{W}$ leads to temporal patterns in $\mathcal{H}$ that are organized according to the underlying rhythm. This effect will be explored in more detail in Section VI.

We have found that imposing sparsity constraints on H_k using an entropic prior [25], [35], ensures more consistent results than using a Dirichlet prior similar to that used for $\mathbf{z}$. This is because, unlike the entropic prior, the hyperparameter of the sparse Dirichlet prior tends to be sensitive to the dimensionality of the underlying distribution. As a result, choosing a single prior setting that works well across many songs is challenging since the dimensionality of $\mathcal{H}$ varies with T. This was not a problem in the previous sections because the dimensions of $\mathcal{W}$ and $\mathbf{z}$ are the same across all songs (assuming a consistent rank K).

An extended discussion about entropic priors for learning structure of multinomial distributions can be found in [35]. For the one-dimensional SI-PLCA model in Section II-C, the prior on $\mathbf{h}_k^T$ can be written as follows:

$$P\left(\mathbf{h}_k^T|\beta_h\right) \propto \exp\left(\beta_h \sum_t h_{kt} \log h_{kt}\right), \quad \beta_h \geq 0. \tag{11}$$

Note that this prior is enforced over the activations for each basis independently, i.e., the rows of $\mathcal{H}$ in Figs. 1 and 3, not on the joint activations across all bases.

In the case of two-dimensional shift-invariance, it is useful to factor $H_k = P(t, r|k)$ into the product of two conditional distributions: $\mathbf{h}_k^T = P(t|k)$, the overall activations as a function of time used in one dimensional SI-PLCA, and $\mathbf{h}_{r|k}^T = P(r|t, k)$, the key modulation of each basis at each point in time. The prior over $\mathbf{h}_{r|k}^T$ has the same form as (11) with an associated hyperparameter β_r.

Sparse activations are especially important in 2-D SI-PLCA with $R = F$ (i.e., allowing all possible key transpositions), in which case H_k has the same dimensionality as V and unconstrained optimization often leads to degenerate solutions where W_k contains a single nonzero element and H_k is an arbitrary transposition of V. The factorization of H_k makes it simple to enforce that only one key rotation be active at any point in time, a necessary constraint for avoiding these degenerate solutions, by setting β_r to a large value.

IV. Parameter Estimation

The parameters of the decomposition of (6) can be computed iteratively using an expectation–maximization (EM) algorithm. It is worth noting that in the 1-D case without priors, the SI-PLCA EM algorithm leads to update rules which are numerically identical to those of NMF based on a Kullback–Leibler divergence cost function. The full derivation of the SI-PLCA EM algorithm and an exploration of its relationship with NMF can be found in [36]. Here, we review it in the context of the 2-D decomposition described in Section II-D using the prior distributions described in Section III.

The joint log probability of V and the model parameters $\theta = \{\mathcal{W}, \mathbf{z}, \mathcal{H}\}$ given the hyperparameters of the prior distributions described in Section III can be written as follows:

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{f,t} v_{ft} \log \hat{v}_{ft} + \sum_{k} (\alpha_z - 1) \log z_k \\ &+ \sum_{f,k,\tau} (\alpha_{w\tau} - 1) \log w_{kf\tau} + \beta_h \sum_{t,k} h_{kt} \log h_{kt} \\ &+ \beta_r \sum_{t,k,r} h_{r|kt} \log h_{r|kt}. \end{aligned} \tag{12}$$

The EM algorithm finds the settings for θ that maximize the posterior probability in (12) by initializing the distributions W_k, $\mathbf{z}$, and H_k randomly and then iteratively performing the expectation and maximization updates given in the following sections until the parameters converge. This algorithm is only guaranteed to converge to a local optimum, so the quality of the factorization depends on the initialization. In our experiments we found that more consistent results are obtained by initializing $\mathbf{z}$ and H_k to be uniform distributions while setting the initial W_k randomly by sampling each entry in the matrix from a uniform distribution and then normalizing.

A. Expectation Step

In the expectation step, the posterior distribution over the hidden variables k, τ, and r is computed for each cell in V. For notational convenience we represent this distribution as a set of matrices $\{P_{k\tau r}\}$ for each setting of k, τ, and r. Each point in the $F \times T$ matrix $P_{k\tau r}$ corresponds to the probability that the corresponding point in V was generated by basis k at time delay τ and relative key transposition r. It can be computed as follows:

$$P_{k\tau r} \propto z_k \overset{r}{\uparrow} \mathbf{w}_{k\tau} \otimes \overset{\rightarrow\tau}{\mathbf{h}}{}_{kr}^{T} \tag{13}$$

where $\otimes$ denotes the outer product. The set of $P_{k\tau r}$ matrices are normalized such that each point in $\sum_{k\tau r} P_{k\tau r}$ is one.

B. Maximization Step

Given the posterior distribution computed in the E-step, the parameters are updated during the maximization step. First, we define the operator

$$\langle x \rangle_{t,\tau} \triangleq \frac{x}{\sum_{t,\tau} x} \tag{14}$$

to normalize x over the dimensions corresponding to t and τ. The parameter updates can then be written as follows:

$$z_k = \left\langle \sum_{\tau,r} \sum_{f,t} V \cdot P_{k\tau r} + \alpha_z - 1 \right\rangle_{k} \tag{15}$$

$$\mathbf{w}_{k\tau} = \left\langle \sum_{r} \sum_{t} \overset{r}{\downarrow} V \cdot \overset{r}{\downarrow} P_{k\tau r} + \alpha_{w\tau} - 1 \right\rangle_{f,\tau} \tag{16}$$

$$\mathbf{h}_k^T = \left\langle \sum_{\tau,r} \sum_{f} \overset{\leftarrow\tau}{V} \cdot \overset{\leftarrow\tau}{P}_{k\tau r} \right\rangle_{t} \tag{17}$$

$$\mathbf{h}_{r|k}^T = \left\langle \sum_{\tau} \sum_{f} \overset{\leftarrow\tau}{V} \cdot \overset{\leftarrow\tau}{P}_{k\tau r} \right\rangle_{r} \tag{18}$$

where $\cdot$ denotes the element-wise matrix product.

Equations (15) and (16) incorporate the Dirichlet prior distributions over the corresponding parameters. Since the entropic prior over H_k is not conjugate to the multinomial distribution, the final setting for $\mathbf{h}_k^T$ and $\mathbf{h}_{r|k}^T$ requires additional computation. We use the fast approximation to the entropic prior described in [37]. Following the update of (17), $\mathbf{h}_k^T$ is refined by iterating the following updates:

$$\gamma_k = \left\langle \mathbf{h}_k^{T\frac{\nu}{\nu-1}} \right\rangle_{t} \tag{19}$$

$$\mathbf{h}_k^T = \left\langle \beta_h \nu \gamma_k + \sum_{\tau,r} \sum_{f} \overset{\leftarrow\tau}{V} \cdot \overset{\leftarrow\tau}{P}_{k\tau r} \right\rangle_{t} \tag{20}$$

where the approximation parameter ν is fixed at 50. The refinement procedure for $\mathbf{h}_{r|k}^T$ follows the same derivation.

Finally, we note that the Dirichlet prior distributions complicate the M-step slightly since computing (15) and (16) directly sometimes results in negative values for the probabilities in $\mathbf{z}$, and W_k. To ensure that the updates result in valid distributions, any negative elements computed during the M-step are clamped to zero.

The overall complexity of a single iteration of the EM algorithm is $O(FTKLR)$, i.e., it scales linearly in the size of the decomposition. It is often possible to significantly speed up the computation by dropping W_k and H_k from the model when z_k falls to zero due to the sparse prior distribution.

In the following sections, we describe how various configurations of the proposed algorithm can be used in a number of music signal processing applications, including the extraction of repeated riffs in popular music, the identification of musical meter, and long-term segmentation into verse-chorus sections.

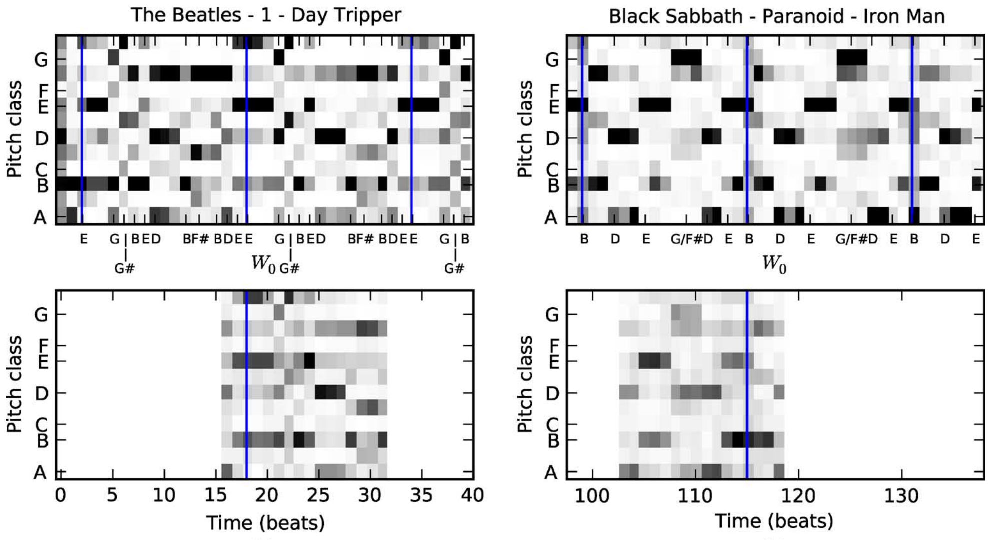
V. Riff Identification

In this section, we describe how the proposed SI-PLCA algorithm can be used to identify repeated motifs in a piece of music. In order to illustrate the ability of the algorithm to identify simple repetitions of a single motif, we address the specific case of riff-driven popular music in which a single chord progression is repeated throughout an entire song.

The analysis described in this paper can naturally identify these repeated riffs using decompositions with a single basis ($K = 1$) and utilizing sparsity constraints to learn the riff length and to control the activation structure. The resulting W_0 pattern corresponds to the most commonly repeated progression in the song and H_0 serves as the high level "score," showing the location of repetitions throughout the song, including any key modulations.

Because the analysis is shift invariant, the key transposition and phase offset of the identified riff are arbitrary and depend on the random initialization. We therefore post-process W_0 and H_0 in order to normalize them to a standard key modulation. We identify this by summing H_k along the columns and selecting the largest value, corresponding to the most often utilized key modulation. We then rotate W_k and H_k to match this modulation index.

Fig. 4 shows example riffs extracted using 2-D SI-PLCA with $R = 12$, allowing for all possible key modulations, maximum



(a) (b)

Fig. 4. Main riffs identified in (a) *Day Tripper* by The Beatles and (b) *Iron Man* by Black Sabbath. The top panels show chromagram excerpts from each song including two repetitions of the main riff. The bottom panels show the identified riff W_0 aligned against the top panels. Blue vertical lines indicate the beginning of the riff.

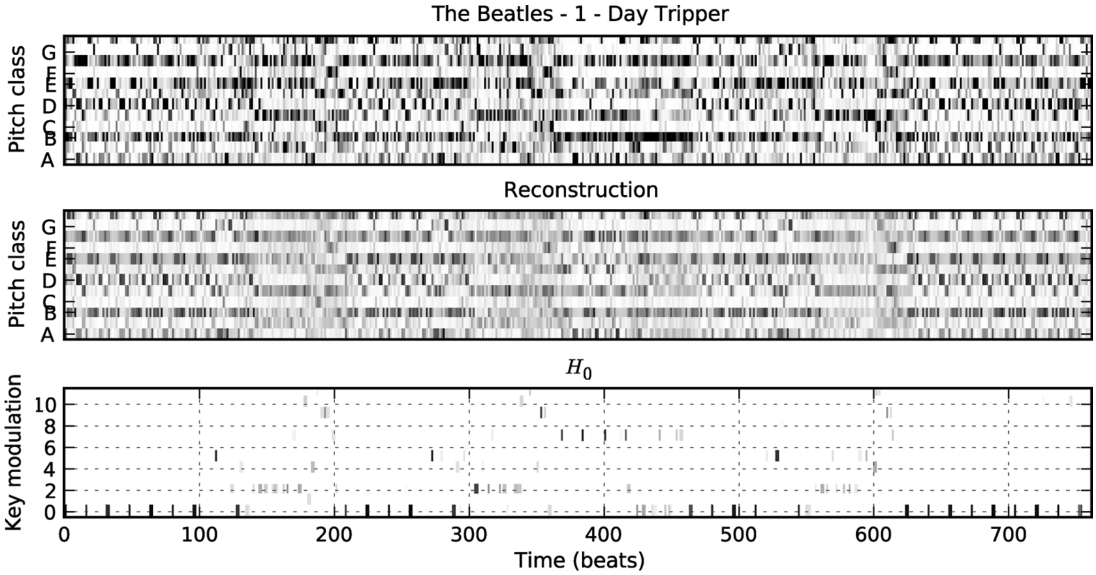


Fig. 5. Activations of the *Day Tripper* riff shown in Fig. 4(a) (bottom panel). Also shown are the original chromagram (top panel) and its reconstruction (middle panel).

basis length $L = 40$, $\alpha_{w\tau}$ constructed using $c = 10$, and $m = -0.0003$ to identify the underlying pattern length, $\beta_r = 1$ to enforce that only one key is active at each point in time, and $\beta_h = 0.1$ to encourage the identification of complete patterns. In both cases the analysis correctly identifies the 4-measure, 16-beat long chord progression that closely matches the contents of V. The key normalization process is accurate as well, as shown by the match between the ground truth transcription and the corresponding patterns. However, the patterns are not properly aligned against the downbeats of the original song. This is because the repetitions of the riff are for the most part periodic, and the analysis does not make use of any phase analysis, resulting in a random phase offset. This problem could be addressed using additional post-processing to align W_0 and H_0 to the downbeat locations in the signal; however, we leave this for future work.

The activations corresponding to Fig. 4(a) are shown in Fig. 5. Throughout most of the song, activations occur every 16 beats, corresponding to the length of the main riff. Because we limit the decomposition to a single basis, it is forced to utilize the main riff to reconstruct sections of the song to which it is not well matched, leading to noisier activations and larger reconstruction error in those sections. Such activation patterns are visible during beats 140–200, 300–360, and 550–610, corresponding to the end of the verse sections which are mainly in the key of F#.

Finally, we note that H_0 correctly transcribes the key modulations used within the piece. The song is in the key of E major, corresponding to key modulation $r = 0$. The main riff is transposed to the subdominant (A, $r = 5$ semitones) during the verse, visible at beats 110, 270, and 530, and to the dominant (B, $r = 7$ semitones) during the bridge, visible between beats 370 and 430.

In order to extend this algorithm to the more general problem of identifying recurring motifs scattered throughout a long piece, a similar anlysis could be used, albeit using a larger value for K. Additional post-processing would be necessary to differentiate between bases that correspond to repeated motifs, and those that are not activated as frequently.

Another application is that of music thumbnailing [17], which seeks to automatically identify a representative excerpt from a song. Utilizing the analysis described in this section, the thumbnail that best represents the identified motif (typically a good thumbnail since it is the most often repeated pattern found in the song) simply corresponds to the largest activation in $\mathcal{H}$. If a longer thumbnail is required, then L or the parameters of $\alpha_{w\tau}$ can be scaled up to identify longer patterns.

VI. Meter Analysis

In this section, we demonstrate how the structure of SI-PLCA activations encodes information about the rhythmic content of a piece of music. However, instead of performing beat or downbeat tracking as in much music informatics research, e.g., [38], we use a simple analysis of $\mathcal{H}$ to discriminate between different metrical patterns, similar to the task in [39].

In most western music, there is a strong relationship between chroma and meter in that chord changes are much more likely to occur on the downbeat than at any other metrical position. This observation has been used for downbeat tracking in [40] and to reinforce chord detection in [41]. Although it is possible that more accurate meter analysis could be obtained by augmenting or replacing the mid-level chroma representation with onset-based features, we choose to focus on this representation in order to emphasize the versatility of the proposed algorithm to many applications with minimal modifications.

Our approach is based on [42], where time signatures are characterized using the autocorrelation of note onset times in the musical score. Likewise, we hypothesize that songs in the same time signatures will show consistent periodicity in their activations, as characterized by the autocorrelation of $\mathcal{H}$. As in Section V, we reconstruct the song's features using a single basis, thus assuming that the time signature remains constant throughout the song. Given the SI-PLCA decomposition with $K = R = 1$, features are computed by taking the 1024 point power spectrum of $\mathbf{h}_0$, encoding equivalent information to the global autocorrelation across the entire song. Example features can be seen in Fig. 6.

In order to evaluate this hypothesis, we constructed a data set of 342 pop songs in various time signatures. The data set was broken into four metrical classes as follows: class 3 contains 144 songs in 3/4, 6/4, 6/8, and 9/8 time (i.e., songs in triple meter), similarly class 4 contains 155 songs in 4/4 time, class 5 contains 25 songs in 5/4 and 10/4 time, and class 7 contains 18 songs in 7/4 and 7/8 time. We evenly split the songs in each class between training and testing sets and use the training set to build a simple classifier based on template-matching. The template for each meter class is found by averaging the power spectrum features over all training songs in the same class. Fig. 6 shows the trained templates for the four meter classes. Each class is clearly characterized by a different harmonic series: the triple meter class (top panel) has its most prominent peaks corresponding to periods of 3, 6, and 12, while the 4/4 class (second panel) has peaks at periods of 2, 4, and 8 beats, etc.

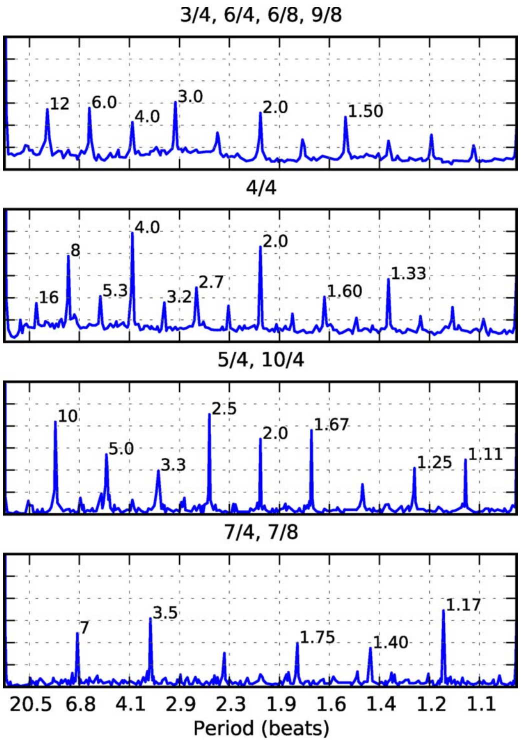


Fig. 6. Meter templates trained over the data set described in Section VI.

Points in the held out test set are classified according to the nearest template using a Euclidean distance metric. A grid search over SI-PLCA parameters was performed to find the best performing features, achieved using $L = 60$, $c = 10$, $m = 5 \times 10^{-6}$, and $\beta_h = 0.1$. Using these features we obtain overall classification accuracy of 61%. The confusion matrix is shown in Table I. Songs in triple meter are most often confused with songs in 4/4, likely due to the peaks at 2 and 4 beats shared between both classes. This shared periodicity is a consequence of the ambiguous nature of the compound 6/4 and 6/8 time signatures, in which measures are sometimes broken into repeated units of 2 and 4 beats instead of 3 and 3. Triple meter also shows significant confusion with 7/4 time, despite not sharing any significant peaks. This is primarily a result of the fact that the 7/4 template is very close to zero in regions between harmonic peaks, a consequence of overfitting to the very small amount of training data available. Finally, class 5 also suffers as a result of overfitting and has the worst overall performance, also showing significant confusion with 4/4 and 7/4.

When restricted to the simpler task of differentiating between duple and triple meter, performance improves to 78%. Although we do not claim that these results are state-of-the-art, especially

TABLE I
METER CLASSIFICATION CONFUSION MATRIX SHOWING THE DISTRIBUTION OF CLASSIFICATION ERRORS IN THE TEST SET. THE ENTRIES CONTAIN RAW SAMPLE COUNTS, NOT PERCENTAGES

		Predicted			
	Class	3	4	5	7
True	3	35	15	8	14
	4	13	59	1	4
	5	0	5	3	4
	7	2	0	1	6

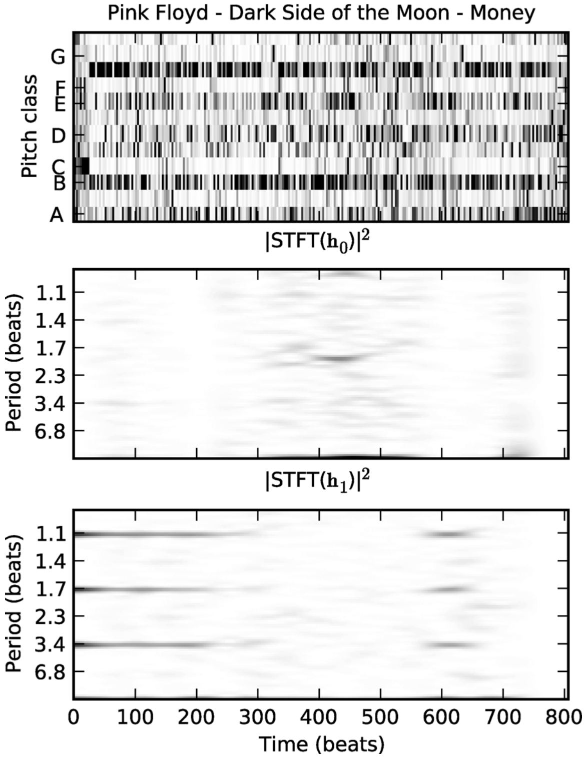


Fig. 7. Meter analysis of *Money* by Pink Floyd using an SI-PLCA decomposition with $K = 2$. The bottom two panels show the short time power spectrum of $\mathbf{h}_0$ and $\mathbf{h}_1$, corresponding to sections in 4/4 and 7/4 time, respectively.

given the simplicity of the classifier, they indicate that the activations in the proposed analysis tend to be very strongly related to a song's underlying metrical structure.

For songs containing meter changes, sections dominated by different time signatures can be identified using an SI-PLCA decomposition with K larger than one. Each component of this decomposition will correspond to a different repeated harmonic pattern, potentially in a different meter.

An example of such a meter change in Pink Floyd's *Money* is shown in Fig. 7. The majority of the song is in 7/4 time, with a shorter section in 4/4. We approximate the chromagram in the top panel of the figure using a rank 2 SI-PLCA decomposition with the same parameter settings described above, and visualize the meter change using the short-time power spectra of the activations $\mathbf{h}_0$ and $\mathbf{h}_1$, shown in the middle and bottom panels, respectively. The spectrum of $\mathbf{h}_1$ features strong spectral lines at periods of 3.5, 1.75, and 1.17 beats, clearly corresponding to peaks in the 7/4, 7/8 template in Fig. 6. Similarly, the spectrum of $\mathbf{h}_0$ has a prominent peak at about 2 beats, indicating duple time, corresponding to the section in 4/4 time. Note that the periodic structure of $\mathbf{h}_0$ is not as well defined as that of $\mathbf{h}_1$ since it corresponds to an extended guitar solo, which, unlike basis 1, is not composed of nearly exact repetitions of a single motif. This leads to noisier activations.

In the following section we continue the discussion of SI-PLCA decompositions with rank larger than one, however, we shift the focus to the identification of long-term temporal structure in music.

VII. STRUCTURE SEGMENTATION

On longer time scales, repetitive patterns in popular music appear as repetitions of entire sections such as verse, chorus, and bridge. Given long enough bases, the analysis described in this paper naturally identifies such long-term temporal structure within a song, encoded by the activations in $\mathcal{H}$.

As before, we use the one-dimensional version of the algorithm (i.e., $\mathrm{R} = 1$), since long-term key modulations often indicate new sections in popular music. We assume a one-to-one mapping between the identified harmonic patterns and the underlying song structure, i.e., we assume that each pattern is used within only one segment. The mapping is derived by computing the contribution of each pattern to the chromagram by summing (6) across all pitch classes:

$$\ell_k(t) = P(t,k) = \sum_f \hat{v}_{kft} \tag{21}$$

$$\hat{V}_k = z_k W_k * \mathbf{h}_k^T. \tag{22}$$

The quantity in (21) corresponds to the probability that the observation at time t comes from basis k. We assume that each basis corresponds to a unique segment label and compute the final segmentation from $\ell_k(t)$ by finding the optimal setting of k at each time frame. We constrain this path through (21) using a simple transition matrix designed to smooth out transitions between segments, and compute the optimal path using the Viterbi algorithm. The transition matrix is constructed to have a large weight along the diagonal to discourage spurious transitions between segments. The off diagonal components are uniform, so no preference is given to any particular state:

$$a_{ij} = \begin{cases} p, & i = j \\ \frac{1}{K-1}(1-p), & i \neq j \end{cases}. \tag{23}$$

p is set to 0.9 throughout the experiments in this section. Finally, the per-frame segment labels are post-processed to remove segments shorter than a predefined minimum segment length.

A. Examples

Fig. 8 shows an example of the segmentation process using the decomposition from Fig. 1. The top panel shows the original chromagram of the song. The following four panels show the contribution of each pattern to the chromagram, and the bottom two panels show $\ell_k(t)$ and the final segmentation, respectively.

There are some interesting differences between the ground truth segmentation and that derived from the proposed algorithm

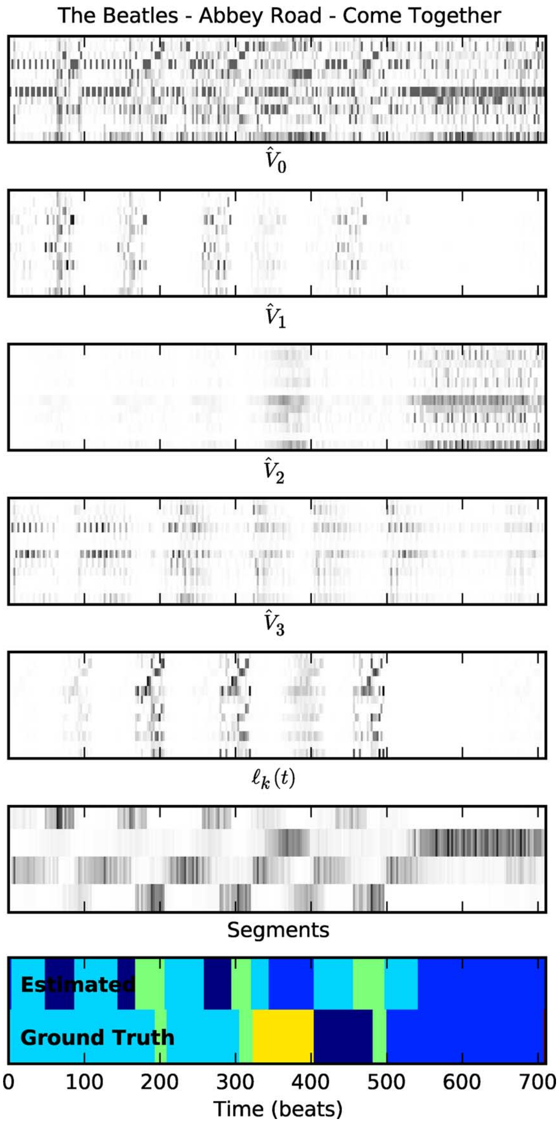


Fig. 8. Song structure segmentation using the SI-PLCA decomposition shown in Fig. 1. The pairwise F-measure of the estimated segmentation is 0.5.

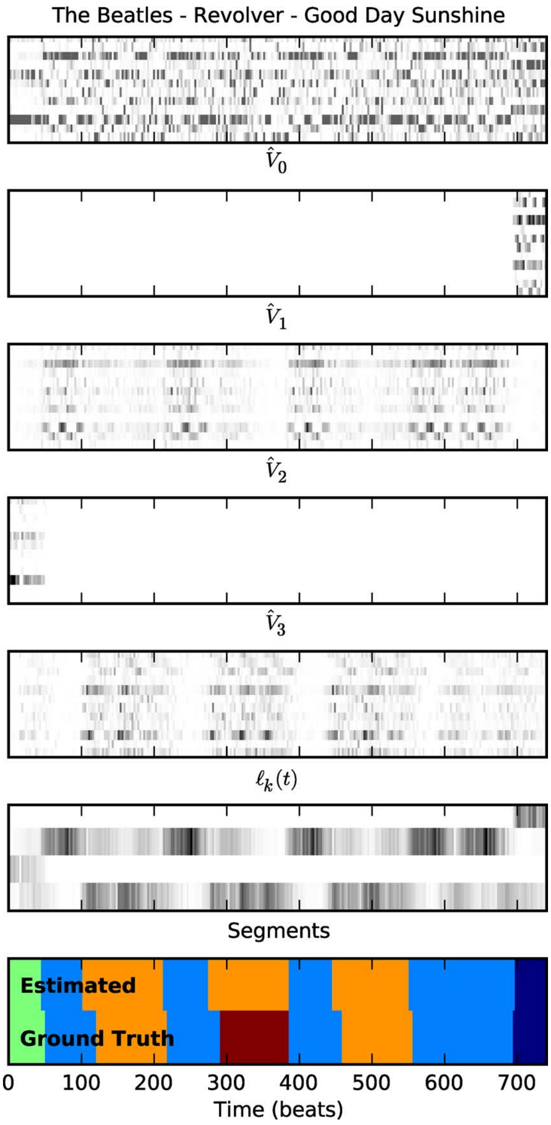


Fig. 9. Song structure segmentation using the SI-PLCA decomposition shown in Fig. 3 $(\mathrm{PFM} = 0.73)$.

in Fig. 8. For example, the proposed algorithm breaks the beginning of the song into repeated subsections: basis 2 (cyan) → basis 0 (dark blue), while the ground truth labels this sequence as a single segment. When inspecting the actual patterns it is clear that these ground truth segments are composed of distinct chord patterns, despite serving a single musical role together ("intro/verse" as annotated in the ground truth). In fact the cyan and dark blue segments are reused in different contexts throughout the song in regions with different ground-truth annotations. The analysis has no notion of musical role, and tends to converge on solutions in which bases are reused as often as possible. This can be considered a limitation of our segmentation algorithm, which could be addressed using more sophisticated post-processing to combine segment labels that often occur together.

Another method of addressing this would be to increase the length L of the convolutive bases (or the corresponding parameters of $\alpha_{w\tau}$), in which case the repeated sub-segments would be merged into a single long segment. This highlights an inherent tradeoff in the proposed analysis between identifying simple chord patterns that are frequently repeated (short W_k, many activations in $\mathbf{h}_k^T$) as opposed to deriving long-term musical structure (longer W_k, sparser $\mathbf{h}_k^T$). This is a recognized ambiguity in the concept of musical segmentation [43].

When high-level segments are more closely correlated with the harmonic structure identified by our method, the proposed analysis leads to good segmentation. An example of this, based on the decomposition shown in Fig. 3, is depicted in Fig. 9. Note that the ground truth labels make a functional distinction between "verse" (orange) and "verse/break" (red) which is not present in our analysis.

B. Experiments

In the following, we evaluate the proposed approach to structure segmentation. We quantify the effect of the various prior distributions described in Section III and compare our approach to other state-of-the-art algorithms. The test set consists of 180 songs from the recorded catalog of The Beatles, annotated into verse, chorus, refrain, etc., sections by the Centre for Digital Music.[1] Each song contains an average of about ten segments

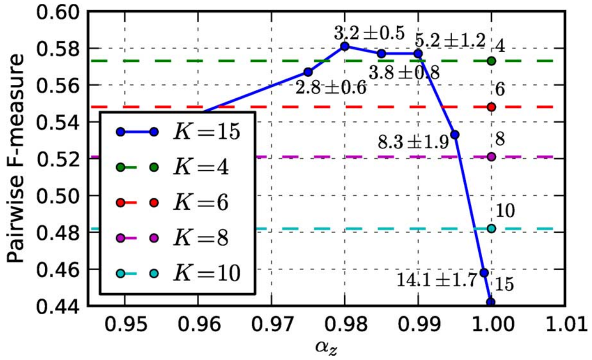


Fig. 10. PFM as a function of α_z (solid line). $K = 15$, $L = 60$, and no other priors are used. The average effective rank and standard deviation are displayed for each setting of α_z. Also plotted is PFM for $\alpha_z = 1$ for different settings of K (dashed lines).

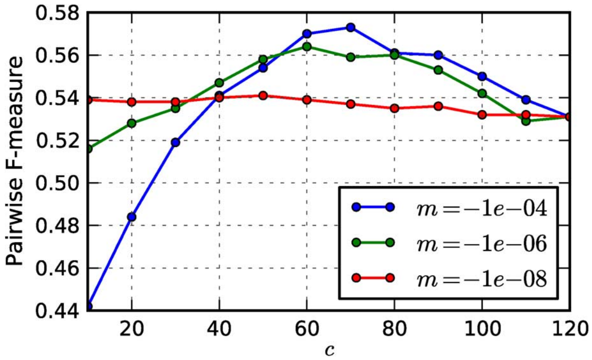


Fig. 11. PFM as a function of the $\alpha_{w\tau}$ cutoff parameter c for varying slopes m. The rank is fixed at $K = 4$ and the maximum basis length is fixed at $L = 120$.

and 5.6 unique labels. Note that all results reported in this section are computed over the entire data set.

Segmentation performance is measured using the pairwise recall rate (PRR), precision rate (PPR), and F-measure (PFM) metrics proposed in [44] which measure the frame-wise agreement between the ground truth and estimated segmentation regardless of the exact segment label. We also report the entropy-based over- and under-segmentation scores (S_o and S_u, respectively) as proposed in [45].

1) Number of Patterns: Since our segmentation algorithm assumes a one-to-one relationship between patterns and segments, the appropriate choice of the number of patterns K is critical to obtaining good performance. We evaluate this effect by segmenting the data set with varying settings for K with $\alpha_z = 1$, and by fixing K to 15 and varying α_z. In all cases, L is set to 60 and no other priors are used.

The results are shown in Fig. 10. For $\alpha_z = 1$, segmentation performance decreases as K increases, peaking at $K = 4$. Performance improves when the sparse prior is applied for most settings of α_z. The average effective rank and its standard deviation both increase with decreasing α_z (increasing sparsity). The best performance is obtained for $\alpha_z = 0.98$, leading to an average effective rank of 3.2 ± 0.5. These results demonstrate the advantage of allowing the number of patterns to adapt to each song.

2) Pattern Length: As described in Section VII-A, the length of the patterns used in the decomposition has a large qualitative effect on the segmentation. To measure this effect, we segmented the entire corpus varying the cutoff parameter c between 10 and 120 beats under different settings of the sparseness penalty m. The results are shown in Fig. 11.

The best performance occurred with small m, corresponding to the use of bases of roughly fixed length c. Fixing $m = 10^{-4}$ and varying c (blue curve) results in poor performance for small c since the ground truth segments are often divided into many distinct short segments. Performance improves with increasing c, until it reaches a peak at $c = 70$. When c grows larger than the average segment length in the ground truth (78 beats) the performance decreases.

Decreasing m to 10^{-6} (green curve), i.e., allowing the basis length more freedom to adapt to particular songs, shows slightly reduced segmentation performance in the best case, but less overall sensitivity to the exact value of the cutoff. The worst performance occurs at the other extreme of m very close to zero (red curve). In this case $\alpha_{w\tau} \approx 1$ for all τ, corresponding to bases of roughly fixed length L for all settings of c.

Eliminating the sparse prior over W_k and varying L leads to nearly identical segmentation performance to using $m \leq 10^{-4}$. We can therefore conclude that there is no advantage to allowing for varying pattern length in this task. Following this trend, we have also found that $\beta_h > 0$ has minimal effect on performance, so it is not used in the remaining experiments. These results are not surprising since the segmentation is derived from the *combination* of $\mathcal{W}$ and $\mathcal{H}$. Shifting the sparsity from one factor to another does not have any significant impact on $\ell_k(t)$.

3) Comparison to the State-of-the-Art: We compare the proposed segmentation system with other state-of-the-art approaches, including Levy and Sandler's HMM-based segmentation system[2] [44] (QMUL) and a more recent system from Mauch *et al.* [46] based on analysis of self-similarity matrices derived from beat-synchronous chroma. As in Section VII-B1, we found that QMUL has optimal PFM when the number of segments is set to 4.

We compare these to the proposed system using fixed rank $K = 4$ (SI-PLCA) and a variant using sparse $\mathbf{z}$ with $\alpha_z = 0.985$ and $K = 15$ (SI-PLCA-α_z). L was fixed at 70 for both systems, and the minimum segment length was set to 32. Also included is a baseline random segmentation in which each song is divided into fixed length segments of 32 beats, and each segment is given one of four randomly selected labels.

The results are shown in Table II. The system from Mauch *et al.* performs best, followed by SI-PLCA-α_z, SI-PLCA, and QMUL. All systems perform significantly better than the baseline. All of the segmentation systems have roughly comparable pairwise precision and S_u. The differences are primarily in the recall (and S_o) with Mauch *et al.* outperforming SI-PLCA-α_z by 8% (14%), and SI-PLCA-α_z in turn outperforming QMUL by 16% (12%). The proposed segmentation algorithm showed similar performance characteristics, in many cases roughly

[1] http://isophonics.net/content/reference-annotations-beatles

[2] Available online: http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html

TABLE II
SEGMENTATION PERFORMANCE ON THE BEATLES DATA SET. THE NUMBER OF LABELS PER SONG WAS FIXED TO 4 FOR SI-PLCA, QMUL, AND RANDOM. THE AVERAGE EFFECTIVE RANKS FOR SI-PLCA-α_z AND MAUCH *ET AL* WERE 3.9 AND 5.5, RESPECTIVELY

System	PFM	PPR	PRR	S_o	S_u
Mauch et al [46]	0.66	0.61	0.77	0.76	0.64
SI-PLCA-α_z	0.60	0.57	0.69	0.62	0.56
SI-PLCA	0.59	0.60	0.61	0.57	0.57
QMUL [44]	0.54	0.58	0.53	0.50	0.57
Random	0.47	0.43	0.56	0.39	0.41

comparable with the other top performers, in the Music Structure Segmentation task of the most recent Music Information Retrieval Evaluation Exchange (MIREX).[3]

The primary shortcoming of the proposed algorithm lies in its tendency to over-segment as seen in the example in Fig. 8. This could be addressed using a set of heuristics to merge segment labels that frequently follow one another, however we leave this for future work. Many of the other qualitative differences in performance between Mauch *et al.* and SI-PLCA-α_z are a result of more accurate boundary detection in the former system, due in part to special care taken to only allow segments to begin at likely measure boundaries. In contrast, the proposed system often has poor alignment to the underlying measure structure, as seen in the examples in Figs. 4 and 9.

VIII. CONCLUSION

We describe an unsupervised algorithm for identifying repeated patterns in music audio using shift-invariant probabilistic latent component analysis. The analysis can be used to extract a temporal structure information across different time scales by varying its parameter settings. We demonstrate that the use of sparse prior distributions over the SI-PLCA parameters can be used to automatically identify the bases that are most relevant for modeling the data and discard those whose contribution is small. We also demonstrate a similar approach to estimating the optimal length of each basis. The use of these prior distributions enables a more flexible analysis and eliminates the need to specify these parameters exactly in advance.

Finally, we show how the approach can be successfully applied to motif finding, meter analysis, and structure segmentation of popular music. In all cases, there is potential for improvement using more sophisticated post-processing mechanisms, e.g., downbeat alignment for riff extraction, a better classifier for meter identification, or clustering of pattern contributions for segmentation, but such additions are beyond the scope of this paper. To encourage the investigation of these and other ideas, we make the source code freely available online.[4]

Beyond these applications, there are many other scenarios where the proposed method can prove useful, both in the analysis of individual songs and music collections. For example, it can be used to search for common motifs throughout a corpus of music as in [47], and applied to the retrieval of cover songs and other musical variations. Similarly, as demonstrated by Mauch *et al.* in [46], knowledge of repeated patterns can be used to improve automatic chord recognition performance, by helping to smooth over feature variations. In the context of the proposed analysis this amounts to simply analyzing the bases W_k instead of independently analyzing each realization of the corresponding motif in the chromagram.

Finally, our main focus for future work will be on extending the algorithm to be invariant to time-warping (i.e., nonlinear shift-invariance), thus allowing the robust identification of all instances of a given pattern despite variations in length. In the context of the feature representation used in this paper, this extension will prove most useful in the presence of beat-tracking inconsistencies. Furthermore, it will free the analysis from requiring an event-synchronous feature representation, by allowing the use of fixed-hop size feature sequences. This may prove important for the analysis of expressive music signals, e.g., classical music, where beats are difficult to track, or non-musical signals, such as speech or environmental sound, which are not naturally aligned to a regular time grid.

ACKNOWLEDGMENT

The authors would like to thank G. Grindlay, P. Smaragdis, and M. Hoffman for many useful discussions about this work, J. Forsyth for his help in assembling the meter data set, and M. Mauch for sharing the implementation of the segmentation algorithm from [46].

REFERENCES

[1] R. Middleton, "Form," in *Key Terms in Popular Music and Culture*, B. Horner and T. Swiss, Eds. New York: Wiley-Blackwell, 1999, pp. 141–155.

[2] R. Rowe*, Machine Musicianship*. Cambridge, MA: MIT Press, 2004.

[3] A. Ockelford*, Repetition in Music: Theoretical and Metatheoretical Perspectives*. Surrey, U.K.: Ashgate, 2005, vol. 13, Royal Musical Association monographs.

[4] J.-L. Hsu, C.-C. Liu, and A. Chen, "Discovering nontrivial repeating patterns in music data," *IEEE Trans. Multimedia*, vol. 3, no. 3, pp. 311–325, Sep. 2001.

[5] E. Cambouropoulos, "Musical parallelism and melodic segmentation: A computational approach," *Music Percept.*, vol. 23, pp. 249–268, 2006.

[6] O. Lartillot, "A musical pattern discovery system founded on a modeling of listening strategies," *Comput. Music J.*, vol. 28, no. 3, pp. 53–67, 2004.

[7] D. Conklin, "Representation and discovery of vertical patterns in music," in *Proc. Int. Conf. Music Artif. Intell. (ICMAI)*, 2002, pp. 32–42.

[8] D. Meredith, K. Lemström, and G. A. Wiggins, "Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music," *J. New Music Res.*, vol. 31, no. 4, pp. 321–345, 2003.

[9] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Aug. 2010, pp. 625–636.

[10] R. B. Dannenberg and N. Hu, "Discovering musical structure in audio recordings," in *Proc. Int. Conf. Music Artif. Intell. (ICMAI)*, 2002, pp. 43–57.

[11] B. Logan and S. Chu, "Music summarization using key phrases," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000, pp. 749–752.

[12] J.-J. Aucouturier and M. Sandler, "Finding repeating patterns in acoustic musical signals," in *Proc. 22nd Int. AES Conf. Virtual, Synth., Entertainment Audio*, 2002.

[13] G. Peeters, A. L. Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2002, pp. 94–100.

[3] http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results

[4] http://marl.smusic.nyu.edu/resources/siplca-segmentation/

[14] S. A. Abdallah, K. Noland, M. B. Sandler, M. Casey, and C. Rhodes, "Theory and evaluation of a Bayesian music structure extractor," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2005, pp. 420–425.

[15] J. Foote, "Visualizing music and audio using self-similarity," in *Proc. 7th ACM Int. Conf. Multimedia*, 1999, pp. 77–80.

[16] J. Foote, "The beat spectrum: A new approach to rhythm analysis," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2001, pp. 881–884.

[17] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 96–104, Feb. 2005.

[18] M. Goto, "A chorus-section detecting method for musical audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. V-437–V-440.

[19] M. Müller*, Information Retrieval for Music and Motion*. New York, NJ: Springer-Verlag, 2007.

[20] J. Paulus and A. Klapuri, "Music structure analysis using a probabilistic fitness measure and a greedy search algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1159–1170, Aug. 2009.

[21] M. Marolt, "A mid-level representation for melody-based retrieval in audio collections," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1617–1625, Dec. 2008.

[22] L. Wang, E. Chng, and H. Li, "A tree-construction search approach for multivariate time series motifs discovery," *Pattern Recognition Lett.*, vol. 31, no. 9, pp. 869–875, 2010.

[23] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2005, pp. 288–295.

[24] M. Casey and M. Slaney, "Song intersection by approximate nearest neighbor search," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006.

[25] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 2069–2072.

[26] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Proc. ISCA Tutorial and Research Workshop Statist. and Percept. Audition (SAPA)*, 2004.

[27] R. J. Weiss and J. P. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Aug. 2010, pp. 123–128.

[28] M. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA)*, 2006, pp. 700–707.

[29] D. Ellis and G. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, pp. IV-1429–IV-1432.

[30] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[31] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," *Adv. Neural Inf. Process. Syst.*, pp. 1313–1320, 2008.

[32] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1–2, pp. 177–196, 2001.

[33] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[34] V. Tan and C. Févotte, "Automatic relevance determination in non-negative matrix factorization," in *Proc. Signal Process. With Adaptive Sparse Structured Represent. (SPARS)*, 2009.

[35] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Comput.*, vol. 11, no. 5, pp. 1155–1182, 1999.

[36] P. Smaragdis and B. Raj, "Shift-invariant probabilistic latent component analysis," MERL, Tech. Rep. TR2007-009, Dec. 2007.

[37] M. D. Hoffman, "Approximate maximum *a posteriori* inference with entropic priors," Sep. 2010 [Online]. Available: http://arxiv.org/abs/1009.5761, Tech. Rep.

[38] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.

[39] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Proc. Audio Eng. Soc. Conv. 114*, Mar. 2003.

[40] G. Peeters and H. Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: Theory and large-scale evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. XXX–XXX, Aug.. 2011.

[41] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1280–1289, Aug. 2010.

[42] J. Brown, "Determination of the meter of musical scores by autocorrelation," *J. Acoust. Soc. Amer.*, vol. 94, no. 4, pp. 1953–1953, Oct. 1993.

[43] G. Peeters and E. Deruty, "Is music structure annotation multi-dimensional? A proposal for robust local music annotation," in *Proc. 3rd Int. Workshop Learn. Semant. Audio Signals*, 2009.

[44] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 318–326, Feb. 2008.

[45] H. Lukashevich, "Towards quantitative measures of evaluating song segmentation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2008.

[46] M. Mauch, K. C. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2009, pp. 231–236.

[47] T. Bertin-Mahieux, R. J. Weiss, and D. P. W. Ellis, "Clustering beat-chroma patterns in a large music database," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Aug. 2010, pp. 111–116.

Ron J. Weiss (M'09) received the Ph.D. degree in electrical engineering from Columbia University, New York, in 2009.

During the Ph.D. degree, he worked in the Laboratory for the Recognition of Speech and Audio, Columbia University. From 2009 to 2010, he was a Postdoctoral Researcher in the Music and Audio Research Laboratory at New York University. He is currently a Software Engineer at Google, Inc., New York. His research interests lie at the intersection between audio signal processing and machine learning, focusing on sound source separation, and music information retrieval.

Juan Pablo Bello (M'06) received the Ph.D. degree in electronic engineering from Queen Mary University of London, London, U.K.

During the Ph.D. degree, he was also a Post-Doctoral Researcher and Technical Manager of the Centre for Digital Music, Queen Mary University. Since 2006, he has been an Assistant Professor of music technology at New York University, and a founding member of its Music and Audio Research Laboratory (MARL). He teaches and researches on the computer-based analysis of audio signals and its applications to music information retrieval, digital audio effects, and interactive music systems.

Dr. Bello is a member of the Society for Music Information Retrieval (ISMIR), and a regular reviewer and contributor to digital signal processing and computer music journals and conferences. His work has been supported by scholarships and grants from Venezuela, the U.K., the E.U., and the U.S., including, more recently, a CAREER award from the National Science Foundation. He is also a researcher and member of the Scientific and Medical Advisory Board of Sourcetone, a music and health start-up.