STRUCTURAL SEGMENTATION WITH THE VARIABLE MARKOV ORACLE AND BOUNDARY ADJUSTMENT

*Cheng-i Wang**

UCSD Music Department
chw160@ucsd.edu

Gautham J. Mysore

Adobe Research
gmysore@adobe.com

ABSTRACT

For the music structure segmentation task, one wants to solve two co-existing but sometimes contradicting problems; find global repetitive/homogenous structures and locate accurate local change points. In this paper we propose two algorithms to address these two problems. The algorithms can independently or jointly be plugged into various existing structural segmentation algorithms to improve their results. One algorithm utilizes the *Variable Markov Oracle*, a suffix automaton for multi-variate time series capable of finding repeating segments, is proposed to obtain a self-similarity matrix which encodes the global repetitive structure of a music piece. The other proposed algorithm is an iterative boundary adjustment algorithm refining boundary locations. The algorithms are evaluated against the Beatles-ISO dataset and achieve comparable performance to state-of-the-art.

Index Terms— music structure segmentation, Variable Markov Oracle, self-similarity matrix

1. INTRODUCTION

Automatically recognizing the segmentation of a music piece is not only a fundamental task in music information retrieval research for music structure analysis, but also leads to the development of efficient music content navigation and exploration applications. Reviews of existing work could be found in [1, 2]. Among various approaches, the self-similarity matrix (SSM) has been the fundamental building block for several existing algorithms. An SSM captures global repetitive structures containing essential information for music segmentation. Matrix decomposition of an SSM is widely adopted in existing work. In [3], non-negative matrix factorization (NMF) is used to decompose an SSM into basis functions representing different structural sections. The NMF idea is extended in [4] with a convexity constraint on the weights during decomposition, which leads to more stable results. In [5], ordinal linear discriminant analysis is used to learn feature representations from the singular value decomposition of the time-lag SSM. Spectral clustering is used in [6] to obtain low-dimensional repetition representations from an SSM. Approaches focus on deriving segmentation boundaries from an SSM are also popular. In [7], a checkerboard-like kernel is applied along the diagonal of the SSM to obtain a novelty curve for segmentation boundaries. Structure features are devised in [8] based on time-lag SSM and segmentation boundaries are inferred from structure features.

Approaches based on matrix decomposition or boundary detection represents two aspects of music segmentation problem; finding global structures and local change points. The two problems also corresponds to the categorization of repetition/homogeneous- and novelty-based approach proposed in [2]. In this work, we propose algorithms to address the two aforementioned problems. The two algorithms can independently or jointly be plugged into various existing segmentation algorithms. For finding global repetitive structures, we present a novel method for obtaining SSMs. The method is based on Variable Markov Oracle (VMO) [9], a suffix automaton capable of symbolizing multi-variate time-series and keeping track of its repeated motifs. Since repeating sub-sequences are essential in music structure analysis, it is natural to experiment with the SSM obtained from the VMO (VMO-SSM) in the music structure segmentation task instead of the SSMs obtained with traditional approaches. Conventionally, an SSM is obtained by exhaustively calculating frame-by-frame pairwise distances. For music segmentation tasks, a binary SSM (recurrence plot) is often desired [5, 6, 8]. Nearest-neighbor criterion is often used to obtain a binary SSM from an SSM, but the number of nearest neighbors chosen for each frame is often determined heuristically. The VMO-SSM is directly in binary form and the reduction from continuous distance values to binary values is done implicitly according to information dynamics [10, 11]. For improving the boundary detection accuracy, we propose an iterative boundary adjustment algorithm to post-process the results from segmentation algorithms. The proposed algorithms are evaluated in the music structure segmentation task with the Beatles ISO dataset [12] against existing approaches based on SSMs.

2. SSM FROM VARIABLE MARKOV ORACLE

To obtain a binary SSM, the common approach is to apply k-nearest neighbor thresholding for each frame. Using k-nearest neighbor thresholding gains scale-invariance on the result binary SSM. With the VMO-SSM, instead of obtaining scale-invariance, the emphasis is on tracing similar trajectories in the metric space drawn by the time series.

The VMO is a suffix automaton capable of reducing a multi-variate time series down to a symbolic sequence but still retains repeating sub-sequences as shown in [9]. The VMO stores the information regarding repeating sub-sequences within a time series via *suffix links*. For each observation at time t of the time series with length T indexed by a VMO, a *suffix link*, $\mathtt{sfx}[t] = k$, is created pointing back in time k to where the longest repeated suffix happened. The suffix links not only contain the information regarding repeating sub-sequences, but also imply a frame-to-frame equivalency between t and k given $\mathtt{sfx}[t] = k$ that leads to symbolization of the time-series. Given the symbolized sequence S by a VMO, a binary SSM (VMO-SSM), $R \in \mathbb{R}^{T \times T}$, could be trivially obtained

*This work was performed while interning at Adobe Research and jointly supported by CREL at UCSD.

via, with $t > k$,

$$R_{tk} = \begin{cases} 1 & \text{if } \texttt{sfx}[t] = k, \\ 0 & \text{otherwise,} \end{cases}$$

and filling the main diagonal line with 1.

The construction and model selection algorithms for VMO are documented in [13, 14]. A visualization of how a VMO-SSM is obtained is shown in Fig. 1. The advantage of using a VMO to create an SSM over the traditional frame-by-frame pair-wise distance approach is that a VMO selectively chooses frames to calculate distances with for each frame based on if common suffices are shared between two frames. The selective behavior leads to a more efficient calculation than the traditional exhaustive manner ($O(T \log T)$ versus $O(T^2)$ [15]) and also keeps track of recurrent motifs within the time series. The other difference of using a VMO for SSM calculation is that the reduction from a multivariate time series to a symbolic sequence utilizes the concept of information dynamics [10, 11] that aims at modeling the evolving information dynamics as the time series unfolds itself. In the case for the VMO, an information theoretic measurement, *Information Rate* (IR) [10], is maximized to determine the threshold for frame selection during suffix assignment. Let $x_1^T = \{x_1, x_2, \ldots, x_T\}$ denote time series x with T observations, $H(x)$ the entropy of x, the definition of IR is

$$IR(x_1^{t-1}, x_t) = H(x_t) - H(x_t|x_1^{t-1}).$$

IR is the mutual information between the present and past observations and is maximized when there is a balance between variations and repetitions in the symbolized signal. The exact algorithms for calculating IR with a VMO is provided in [13] and not repeated here.

3. SEGMENTATION ALGORITHM

To show how the VMO-SSM could help improving the music structure segmentation task, and to highlight the difference between the using a VMO-SSM and a traditional SSM, two existing work utilizing binary SSMs are adopted in this work. For both work, their original SSMs are replaced by VMO-SSMs. The first segmentation algorithm is the spectral clustering (SC) approach proposed in [6] and the second is the combination of structure features and segment similarity (SF) proposed in [8].

3.1. Spectral Clustering

In [6], the observation is that the partition of the graph defined by a connectivity matrix into m connected components by spectral clustering is effectively the same as segmenting the time series with m distinguished sections (the total number of segments could be more than m with repetition of any of the sections). A series of operations are applied on the SSM to obtain a connectivity matrix, then spectral clustering is applied on the connectivity matrix to obtain a low-dimensional representation of repetitive structures. The operations include nearest neighbor thresholding, smoothing with a median filter, adding local linkages, balancing local and global linkages, linkage weighting and feature fusion.

By replacing traditional SSM originally used in [6] with the binary VMO-SSM described in section 2, only median filtering and adding local linkages are needed to obtain the connectivity matrix R^+ in this work. The median filter is applied in the diagonal direction to suppress erroneous entries, fill missing blanks and keep sharping edges of the diagonal stripes in the binary SSM

$$R' = \text{median}(R_{i+t,j+t} | t \in -\omega, -\omega+1, \ldots, \omega).$$

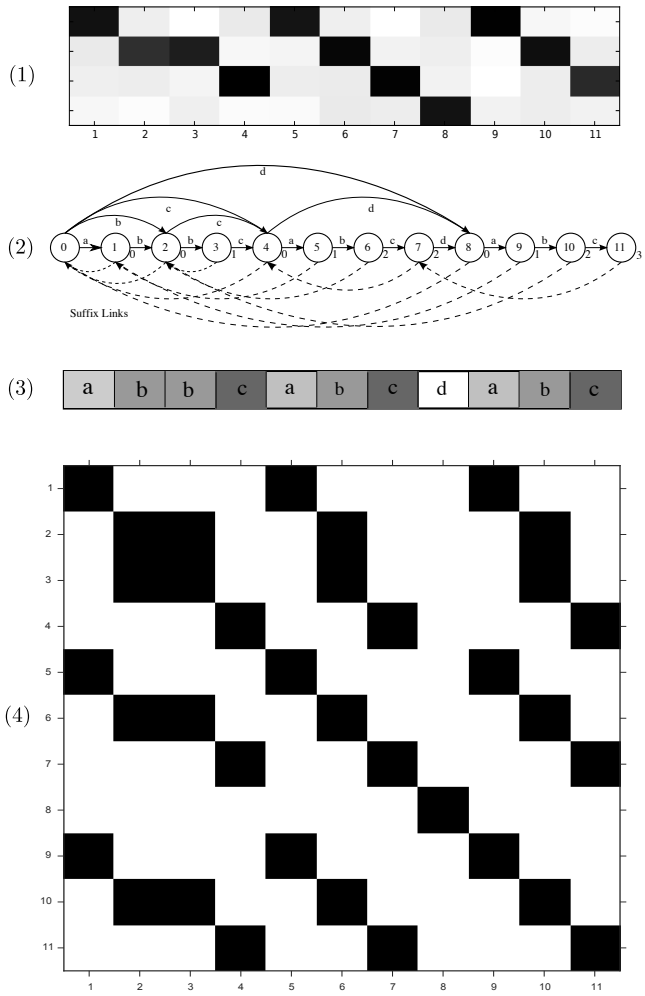

Fig. 1. (1) A synthetic 4-dimensional time series. (2) A VMO structure with symbolized signal $\{a, b, b, c, a, b, c, d, a, b, c\}$, lower (dashed) are suffix links. Values outside of each circle are the length of longest repeated suffix. (3) Symbolized signal. (4) The VMO-SSM obtained from the symbolized signal in (3).

The operation of adding local linkage is defined as

$$\delta_{ij} = \begin{cases} 1 & \text{if } |i-j| = 1, \\ 0 & \text{otherwise} \end{cases}$$

$$R^+_{ij} = \max(\delta_{ij}, R'_{ij}).$$

Let I denote a identify matrix with dimension N, and D the diagonal degree matrix of R^+. The symmetric normalized Laplacian matrix of R^+ is then calculated as

$$L = I - D^{\frac{-1}{2}} R^+ D^{\frac{-1}{2}}.$$

The eigenvectors of L could be interpreted as component membership functions of connected components on a graph defined by L [16]. The segmentation then follows standard spectral clustering algorithm as documented in [16]. In short, the first m eigenvectors with m smallest eigenvalues are concatenated to form a matrix

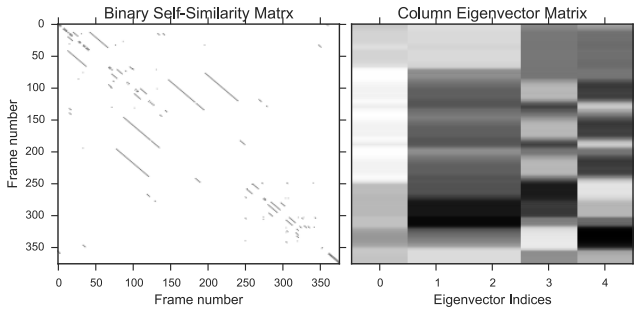

Fig. 2. (Left) The binary VMO-SSM. (Right) The eigenvector matrix, Y, of "All You Need is Love" from Beatles.

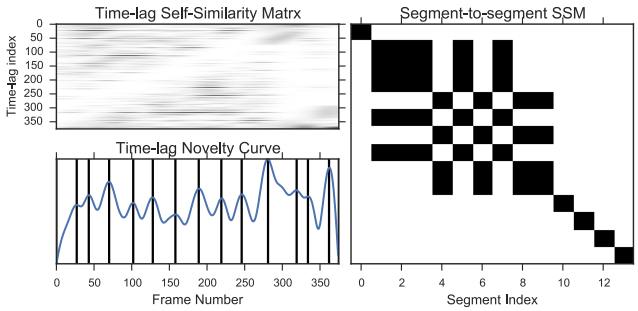

Fig. 3. (Left top) The smoothed time-lag matrix L from a VMO-SSM-SSM. (Left bottom) The novelty curve derived from the time-lag matrix. (Right) The segment-to-segment similarity matrix $\hat{S}$ of "All You Need is Love" from Beatles.

$Y \in \mathbb{R}^{T \times m}$ with rows normalized, then each row of Y is treated as one observation in k-means clustering with $k = m$. The assigned label from k-means clustering is the resulting segmentation labels. Boundaries are inferred from finding label changes between adjacent frames. Visualizations of the R^+ matrix and Y matrix are depicted in Fig. 2.

3.2. Structure Features and Segment Similarity

The details of the SF algorithm could be found in [8]. The goal of that work is to base the algorithm on a local presentation (frame-wise) of global structures (from tim-lag SSM). Similar to the replacement process described in section 3.1, the original binary SSM is replaced by the VMO-SSM. After obtaining R from the VMO, the following steps are applied to find the boundaries first; 1) a time-lag matrix L is obtained from R. 2) L is convolved with a 2-D Gaussian kernel. 3) Boundaries are identified via peak-picking on a novelty curve derived from L. To further obtain segment labels, segment-to-semgent similarities are calculated based on a DTW-like score given R. The resulting similarities are stored in a square matrix $\hat{S}$ with dimensions equal to the number of segments identified from boundary detection. A dynamic threshold based on the statistics of $\hat{S}$ is used to discard non-similar segments. Transitivity between similar segments is induced by iteratively applying matrix multiplication of $\hat{S}$ with itself and threshold. Segment labels are then obtained from the rows of $\hat{S}$. Parameters for this algorithm include the standard deviations of the Gaussian kernel, $\{s_L, s_T\}$, for time-lag and time axis respectively, and peak-picking window length λ. An illustration of L, the novelty curve and $\hat{S}$ derived from R is shown in Fig. 3.

4. BOUNDARY ADJUSTMENT

Observations after examining the segmentation results from last sections reveal that often times the segmentation algorithm is capable of locating the boundaries between segments within a window of a few seconds but is not capable of locating the major change point within a window less than 1 second. The reason might be due to the smoothing on the SSM to obtain R' or L. To remedy the aforementioned situation, an iterative boundary adjustment algorithm is proposed to fine-tune the segmentation boundaries to nearby local maxima in terms of the dissimilarity between adjacent segments.

The criteria to refine boundaries is that the distance between two adjacent segments should be the farthest at the refined boundary points. Based on the criteria, the proposed algorithm adopts the method proposed in [17] where the distance between two segments

Algorithm 1 Iterative Boundary Adjustment

Require: Boundary points B (without beginning and ending frame), features X, window size W, iteration limit N and adjustment cost C.

$n \leftarrow 0$
while True **do**
  $c \leftarrow 0$
  $B' \leftarrow B$
  Randomly permute B'
  for $b \in B'$ **do**
    $\kappa \leftarrow$ K-L divergence of the two segments in X adjacent to b
    $b' \leftarrow b$
    for $t \in \{b - W : b + W\}$ **do**
      $\kappa' \leftarrow$ K-L divergence of the two segments in X adjacent to t
      if $\kappa' > \kappa$ **then**
        $\kappa \leftarrow \kappa$‘
        $b' \leftarrow t$
      end if
    end for
    $b \leftarrow b'$
    $c \mathrel{+}= \text{abs}(b - b')$
  end for
  $B \leftarrow B'$
  $n \mathrel{+}= 1$
  if $c \leq C || n \geq N$ **then**
    break
  end if
end while
return B

are defined as the distance between the empirical distributions of the two segments. The distance criteria boils down to calculating the Kullback-Leibler divergence between the two segments, where the two segments are each modeled by a Multinomial distribution. Since the effect of changing one boundary point propagates to other adjacent segments of neighboring boundaries, an iterative algorithm is devised as shown in Alg. 1.

Alg. 1 resembles an expectation-maximization algorithm in the sense that each iteration (outer for-loop in Alg. 1) stochastically cycles through all boundaries and adjusts them to maximize the K-L divergence of adjacent segments, then fixes the adjusted boundaries as new boundaries and proceeds to the next iteration until convergence criteria are met. The stopping criteria are the total number of iteration N and the total length of boundaries moved C. Empirical observation of running Alg. 1 shows that the total length of boundary moved at each iteration, c, monotonically decreases with number of iterations i.

	Boundaries						Segmentations					
Algorithm	$F_{0.5}$	$P_{0.5}$	$R_{0.5}$	F_3	P_3	R_3	F_{pair}	P_{pair}	R_{pair}	S_f	S_o	S_u
SF (Chroma) [8]	–	–	–	**77.4**	75.3	81.6	**71.1**	78.7	68.1	–	–	–
VMO+SF (Chroma)	36.29	33.84	40.81	69.02	64.27	77.7	61.22	69.99	58.59	**67.38**	64.59	73.25
VMO+SF* (Chroma)	37.37	35.08	41.94	61.5	57.74	68.94	56.16	63.24	54.4	62.81	60.99	67.5
VMO+SC (CQT+MFCC)	34.34	29.38	43.52	64.46	55.09	81.64	55.9	68.63	49.87	62.50	57.59	70.54
VMO+SC* (CQT+MFCC)	**38.41**	34.28	45.47	60.98	54.29	72.26	52.84	61.08	49.05	60.02	55.87	64.84
VMO+SC (Chroma)	31.87	26.39	42.18	61.98	51.2	82.2	52.81	64.57	47.25	59.56	54.93	67.23
VMO+SC* (Chroma)	33.80	28.88	42.07	60.83	52.06	75.45	49.98	57.54	46.40	56.9	53.04	61.37
SC [6] (CQT+MFCC)	31.9	26.03	45.39	57.46	46.95	81.05	54	65.16	48.93	59.56	55.05	67.41
C-NMF [4] (Chroma)	24.89	24.52	26.41	60.41	59.84	63.45	53.53	58.29	52.65	57.2	55.85	60.63
OLDA [5] (Multi-feature)	29.6	29.7	32	53.5	55.3	55	–	–	–	–	–	–
SI-PLCA [18] (Chroma)	28.27	39.57	22.74	50.12	70.59	39.97	49.36	42.67	65.17	48.08	62.28	42.67
CC [19] (Chroma)	25.06	27.3	23.86	55.06	60.17	52.16	49.18	62.91	41.06	56.5	50.36	66.5

Table 1. F, P and R represent F_1-score, precision and recall respectively. Underscores of 0.5 and 3 represent 0.5 and 3 seconds window hit rate scores. $pair$ stands for frame clustering. S_o, S_u and S_f are the normalized conditional entropies of over-, under-segmentation and their F_1-scores. Results of SF and OLDA are from [5, 8]. Results of other algorithms are from [4]. Algorithms followed by (*) are the ones with boundary adjustment algorithm. Parenthesis refers to the feature used for that algorithm. Numbers in bold are the highest F_1-score for each metric.

5. MUSIC STRUCTURE SEGMENTATION

The Beatles-ISO dataset has 179 annotated songs and is widely used in evaluating segmentation algorithms [3,5,8,18,19]. The segmentation experiment aims at identifying a segmentation of a given audio recording and compare the segmentation with human annotations.

5.1. Experiments

To evaluate the effect of the VMO-SSM and the boundary adjustment algorithm, the proposed framework is evaluated against the Beatles-ISO dataset and compared to existing work on the same dataset. Three standard features and their combinations are considered in this experiment; constant-Q transformed spectra (CQT), chroma and MFCCs. All audio recordings are down-sampled to 22050Hz, analyzed with a 93ms window and 23ms hop. CQT are calculated between frequency range $[0, 2093]$Hz with 84 bins. Chroma are derived from CQT by folding the 8 octaves into 12 bins. MFCCs are calculated from 128 Mel bands and 12 MFCCs are taken. All features are beat-synchronized using a beat-tracker [20] with median-aggregation. Similar to [6, 8], features are then stacked using time-delay embedding with one step of history and one step of future. Each dimension of each feature is normalized along the time axis. To combine different features, they are simply stacked and different dimensions are assumed to have equal importance.

A parameter sweep is done to find the best set of parameters in this experiment. Cosine distance is used in the VMO distance calculation. For SC, the median filtering window ω is 17. The number of basis in SC (or the number of distinguished sections), m, is 5. For SF, the standard deviations for time-lag and time axis, $\{s_L, s_T\}$, are 0.5 and 12. The peak-picking window length λ is 9. The parameters for the boundary adjustment algorithm, W, N and C, are $\{4, 10, 2\}$ respectively.

5.2. Evaluation

The evaluation results of the proposed framework along with the ones from other existing work are shown in Table 1. The metrics used follow the ones proposed in Music Information Retrieval Evaluation eXchange (MIREX). The evaluation could be understood in two aspects. The first aspect is the performance on retrieving boundaries and the second one is the performance on assigning labels to regions defined by retrieved boundaries. For boundary hit rate, the combination of the VMO, SC and boundary adjustment outperforms all other existing work by a margin of at least 7% (in [8] it is not reported) for 0.5s window tolerance. For 3s window tolerance, despite being inferior to SF, the approaches using the VMO-SSM are still superior to other existing work. The boundary adjustment algorithm turns out introducing a trade-off between short- and long-time tolerance boundary hit rate. For SC the trade-off of $F_{0.5}$ and F_3 is acceptable with $F_{0.5}$ always improved slightly more than the degradation of F_3. It could be observed that it is not worthwhile applying the boundary adjustment algorithm on SF since the degradation of F_3 is far more than the improvement on $F_{0.5}$. The discrepancy between applying the boundary adjustment algorithm on SC and SF could be understood by the nature of the segmentation algorithms, since SF focuses on finding boundaries from SSM more directly than the matrix decomposition approaches, there might be less room left for improving boundary accuracies in the post-processing stage for SF. For segmentations, original SF ranks the highest in pair-wise clustering F-score and the combination of the VMO and SF is the runner-up. For the F-score of normalized conditional entropy, the VMO-SF combination returns the highest score (it is not reported in [8]). For matrix decomposition approaches, replacing traditional SSMs with VMO-SSMs achieves comparable or superior performances than existing work in segment labeling evaluation.

6. CONCLUSIONS AND DISCUSSIONS

In this work, an alternative SSM extracted from the VMO is shown to be reliable replacing the traditional SSM in music segmentation tasks. In general, using the VMO-SSM improves boundary detection accuracy and achieves comparable or superior performances in segment labeling to state-of-the-art. The reason that the VMO-SSM is better in boundary detection for matrix decomposition approaches is that the selective mechanism during the construction of the VMO suffix structure discards unnecessary calculations, and in turn leads to a cleaner binary SSM than the one from traditional approach.

7. REFERENCES

[1] Roger B Dannenberg and Masataka Goto, "Music structure analysis from acoustic signals," in *Handbook of Signal Processing in Acoustics*, pp. 305–331. Springer, 2008.

[2] Jouni Paulus, Meinard Müller, and Anssi Klapuri, "State of the art report: Audio-based music structure analysis.," in *ISMIR*, 2010, pp. 625–636.

[3] Florian Kaiser and Thomas Sikora, "Music structure discovery in popular music using non-negative matrix factorization.," in *ISMIR*, 2010, pp. 429–434.

[4] Oriol Nieto, *Discovering Structure in Music: Automatic Approaches and Perceptual Evaluations*, Ph.D. thesis, NYU, 2015.

[5] Brian McFee and Daniel PW Ellis, "Learning to segment songs with ordinal linear discriminant analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5197–5201.

[6] Brian McFee and Daniel PW Ellis, "Analyzing song structure with spectral clustering," in *The 15th International Society for Music Information Retrieval Conference*, 2014, pp. 405–410.

[7] Jonathan Foote, "Automatic audio segmentation using a measure of audio novelty," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*. IEEE, 2000, vol. 1, pp. 452–455.

[8] Jean Serra, Mathias Muller, Peter Grosche, and Josep Ll Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *Multimedia, IEEE Transactions on*, vol. 16, no. 5, pp. 1229–1240, 2014.

[9] Cheng-i Wang and Shlomo Dubnov, "Pattern discovery from audio recordings by variable markov oracle: A music information dynamics approach," in *Acoustics, Speech, and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

[10] Shlomo Dubnov, "Spectral anticipations," *Computer Music Journal*, vol. 30, no. 2, pp. 63–83, 2006.

[11] Samer Abdallah and Mark Plumbley, "Information dynamics: patterns of expectation and surprise in the perception of music," *Connection Science*, vol. 21, no. 2-3, pp. 89–117, 2009.

[12] Christopher Harte, *Towards automatic extraction of harmony information from music signals*, Ph.D. thesis, Department of Electronic Engineering, Queen Mary, University of London, 2010.

[13] Shlomo Dubnov, Gérard Assayag, and Arshia Cont, "Audio oracle analysis of musical information rate," in *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*. IEEE, 2011, pp. 567–571.

[14] Cheng-i Wang and Shlomo Dubnov, "The variable markov oracle: Algorithms for human gesture applications," *IEEE MultiMedia*, , no. 1, pp. 54–67, 2015.

[15] Cyril Allauzen, Maxime Crochemore, and Mathieu Raffinot, "Factor oracle: A new structure for pattern matching," in *SOFSEM99: Theory and Practice of Informatics*. Springer, 1999, pp. 295–310.

[16] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[17] Arnaud Dessein and Arshia Cont, "An information-geometric approach to real-time audio segmentation," *Signal Processing Letters, IEEE*, vol. 20, no. 4, pp. 331–334, 2013.

[18] Ron J Weiss and Juan P Bello, "Unsupervised discovery of temporal structure in music," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 6, pp. 1240–1251, 2011.

[19] Mark Levy and Mark Sandler, "Structural segmentation of musical audio by constrained clustering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 318–326, 2008.

[20] Daniel PW Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.