PATTERN DISCOVERY FROM AUDIO RECORDINGS BY VARIABLE MARKOV ORACLE: A MUSIC INFORMATION DYNAMICS APPROACH

Cheng-i Wang, Shlomo Dubnov

University of California, San Diego
CREL, Music Department
{chw160, sdubnov}@ucsd.edu

ABSTRACT

In this paper, a framework for automatic pattern discovery within an audio recording is proposed. The concept of the proposed framework stems from music information dynamics and is realized by *Variable Markov Oracle*. Music information dynamics is the research area focusing on information theoretic measures describing musical structure and is thus closely related to the field of music pattern discovery. *Variable Markov Oracle* is a data structure that provides both fast retrieval of repeated sub-clips from a signal and efficient calculation of music information dynamics measures. Evaluation of the proposed framework is performed on the JKU Patterns Development Dataset with significantly improved performance of the current state of the art.

Index Terms— Music information retrieval, Pattern analysis, Data structures, Variable Markov Oracle

1. INTRODUCTION

Discovering musical patterns (motifs, themes, sections, etc.) is a task defined as identifying salient musical ideas that repeat at least once within a piece [1, 2]. These patterns could potentially be overlapping with each other and not covering the whole piece in contrast to the "segments" found by music segmentation task [3]. In addition, the occurrences of these patterns could be inexact in terms of harmonization, rhythmic pattern, melodic contours, etc. Discovering patterns in musical pieces of either symbolic or audio representations has been investigated [2] and is of interest to the broad community of both music and signal processing. In this paper, the focus is on pattern discovery from audio recordings. For a comprehensive review on studies of symbolic representations, the readers are referred to [2]. Previous researches on pattern discovery from audio recordings either used $F0$-estimation with beat-tracking techniques to enable geometric representation methods on audio recordings [4], or extended music segmentation techniques with greedy search algorithms [5, 6].

In this paper, the use of *Variable Markov Oracle* (*VMO* hereafter) as the basis of the proposed framework differs from other approaches in the sense that music information dynamics is used to identify the musical structure for the pattern discovery task. Research in music information dynamics [7–10] focuses on quantifying the inherent structure of music signals from an information theoretic perspective. Measurements derived from music information dynamics are used in [11–13] as indicators of significant structural changes in the applications of music structure analysis and melody phrase identification, which are closely related to the task described in this paper.

VMO is a data structure capable of symbolizing a signal by clustering the feature frames in the signal, derived from *Factor Oracle* (*FO* hereafter) [14] and *Audio Oracle* (*AO* hereafter) [15]. *FO* is a variant of suffix tree devised for retrieval of patterns from a symbolic sequence [14]. *AO* is the signal extension of *FO* capable of indexing repeated sub-clips of a signal sampled at discrete time, and has been applied to audio query [16], audio structure discovery [13] and machine improvisation [17]. *VMO* was first proposed in [18, 19] for devising an efficient audio query-matching algorithm. In this paper, the capability of using *VMO* to find repeated sub-clips in a signal with an unsupervised manner is shown.

This paper is structured as follows: in section 2, the method of utilizing music information dynamics with *VMO* for music structure / pattern discovery is described; the specifications of the repeated theme discovery task from audio recordings are presented in section 3; evaluation metrics and results are presented in section 4, then conclusions are discussed in section 5.

2. MUSIC INFORMATION DYNAMICS AND *VMO*

Music information dynamics uses information theoretic measures to quantify temporal / structural changes while a music piece unfolds itself [7]. The measurement is achieved by tracing the mutual information between present and past observations of an assumed source random process generating the music signal. Since in most cases the true source process of the music signal generation is unknown, the infor-

Thanks to the Center for Entertainment and Learning (CREL) at UCSD for funding this work.

mation theoretic measures are approximated bsuffixesy either tractable parameterized probabilistic models [8, 9] or compression algorithms [13]. In this paper, the later approach is used since the information theoretic measurements of a music signal could be approximated efficiently by the use of *VMO* via its accompanying compression algorithm [13,20].

2.1. Variable Markov Oracle

VMO symbolizes a signal O, sampled at time t, into a symbolic sequence $Q = q_1, q_2, \ldots, q_t, \ldots, q_T$, with T states and with frame $O[t]$ labeled by a symbol q_t. The symbols are formed by tracking suffix links along the states in an oracle structure. An oracle structure (either *FO*, *AO* or *VMO*) carries three kinds of links, forward link, suffix link and reverse suffix link. Suffix link is a backward pointer that links state t to k with $t > k$, without a label and is denoted by $\texttt{sfx}[t] = k$.

$$\texttt{sfx}[t] = k \iff \text{the longest repeated suffix of } \{q_1, q_2, \ldots, q_t\} \text{ is recognized in } k.$$

Suffix links are used to find repeated patterns in Q. In order to track the longest repeated suffix at each time index t, the length of the longest repeated suffix at each state t is computed by the algorithm described in [14] and is denoted by $\texttt{lrs}[t]$. $\texttt{lrs}$ is essential to the on-line construction algorithm of an oracle structure [14] and its model selection [13]. Reverse suffix link, $\texttt{rsfx}[k] = t$, is basically the suffix link in reverse direction. $\texttt{rsfx}$ is also part of the on-line oracle construction algorithm. $\texttt{sfx}$, $\texttt{lrs}$ and $\texttt{rsfx}$ allow the design of the proposed pattern discovery algorithm described in section 3.2.

Forward links are links with labels and are used to retrieve any of the factors from Q, starting from the beginning of Q and following the path formed by forward links. Since the forward links are not used in the algorithm proposed in this paper, the specifications of it are omitted here and readers are referred to [14] for details.

The last piece needed for the construction of *VMO* is the threshold value, θ. θ is used to determine if the incoming $O[t]$ is similar to to one of the frames following the suffix link started at $t-1$. Two frames, $O[i]$ and $O[j]$, are assigned the same symbol if $|O[i] - O[j]| \leq \theta$. In extreme cases, θ being too low leads to *VMO* assigning different symbols to every frame in O and θ being too high leads to *VMO* assigning the same symbol to every frame in O. As a result, both extreme cases are incapable of capturing any patterns (repeated suffixes) of the signal. In section 2.2, the use of *Information Rate* (*IR* hereafter) to select the optimal θ in the context of music information dynamics is described.

The on-line construction algorithms of *VMO* are proposed in [18] and not repeated here. An example of a constructed *VMO* structure and how $\texttt{lrs}$ and $\texttt{sfx}$ are related to repeated patterns is shown in Fig. 1. The symbols formed by gathering states connected by suffix links have the following properties; 1) states connected by suffix links are guaranteed to have distances less than θ, 2) symbols related to each other sequentially because frames labeled by the same symbol share similar context by the use of suffix links, 3) each state is labeled by one symbol since each state has only one suffix link, 4) the alphabet size of the created symbols is not specified before the construction and is related to the threshold θ value. The suffix structures created by *VMO* or *AO* on the same signal are identical. The advantage of using *VMO* is the computational efficiency provided by the explicit symbolization of the signal done during its construction as described in [18].

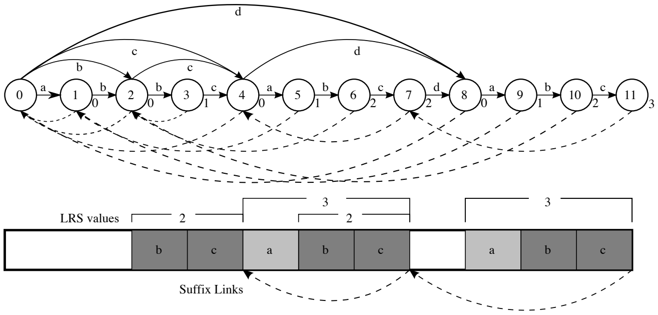

Fig. 1. (Top) A *VMO* structure with symbolized signal $\{a, b, b, c, a, b, c, d, a, b, c\}$, upper (normal) arrows represent forward links with symbols for each frame and lower (dashed) are suffix links. Values outside of each circle are the $\texttt{lrs}$ value for each state. (Bottom) A visualization of how patterns $\{a, b, c\}$ and $\{b, c\}$ are related to lrs and sfx.

2.2. Model Selection via Information Rate

With different θ values, *VMO* constructs different suffix structures and different symbolized sequences from the signal. To select the one sequence with the most informative patterns, ***IR*** is used as the criterion in model selection between the different structures generated by different θ values. ***IR*** is an information theoretic measure capable of measuring the information content of a time series [7] in terms of the predictability of its source process on the present observation given past ones. In the context of pattern discovery with *VMO*, the *VMO* with higher ***IR*** value indicates more of the repeating subsequences (ex. patterns, motives, themes, gestures, etc) are captured than the ones with lower IR value.

Since *VMO* is derived from *AO* [18], the same approach to choose θ by calculating IR is applied here [13]. In brief, given the definition of *IR* and let $x_1^N = \{x_1, x_2, \ldots, x_N\}$ denoting time series x with N observations, $H(x)$ the entropy of x,

$$IR(x_1^{n-1}, x_n) = H(x_n) - H(x_n|x_1^{n-1}), \tag{1}$$

In a nutshell, ***IR*** is the mutual information between the present and past observations, which is maximized when there is a balance between variation and repetition in the symbolized signal. The value of ***IR*** could be approximated by replacing

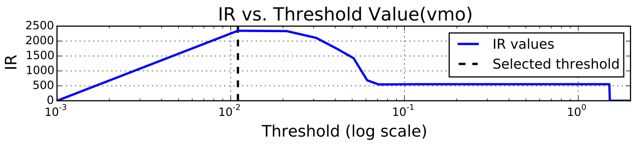

Fig. 2. *IR* values are shown on vertical axis while θ are on horizontal axis. The solid curve in blue color shows the relations between the two quantities and the dashed black line indicates the chosen θ by locating the maximal *IR* value. Empirically the *IR* curves possess quasi-concave function shapes thus global maximum could be located.

the entropy terms in Eq. 1 with complexity measures associated with a compression algorithm, which are the number of bits used to compress x_n independently and compress x_n using the past observations x_1^{n-1}. In [20], a lossless compression algorithm, *Compror*, proven to have similar performance to *gzip* and *bzip2* based on *FO* and `lrs` is provided, and the detail formulation of how *Compror*, *AO* and *IR* are combined is provided in [13]. A visualization of the sum of *IR* values versus different θs on one of the music recordings used in the experiment is depicted in Fig. 2.

3. EXPERIMENT

The dataset chosen for the pattern (repeated theme) discovery experiment is the JKU Pattern Development Dataset [1]. This dataset consists of five polyphonic classical music pieces or movements in both symbolic representation and audio recordings. Ground truth of repeated patterns (themes) for each piece is annotated by multiple musicologists and experts. In this paper, the focus is on repeated pattern discovery from audio recordings. [1]

3.1. Feature Extraction

For the repeated themes discovery task, the feature has to meet the following requirements: 1) it extracts harmonic content of the audio signal in terms of classical Western music tuning; 2) it shows information on a music metrical resolution not at analysis frame level; 3) invariance for the motif with and without harmonization (accompaniment following tonality and chord progression); 4) transposition (moving themes up or down by a constant pitch interval) invariance. Motivated by these four requirements, the feature extraction process is designed as follows. To meet 1), chromagram, a commonly used feature characterizing harmonic content [21], is chosen. For a mono audio recording sampled at 44.1k Hz, the recording is firstly downsampled to 11025Hz, secondly a spectrogram is calculated using a Hann window of length 8192 with 64 samples overlap, then the constant-Q transform of the spectrogram is calculated with frequency analysis ranging between $f_{min} = 27.5$Hz to $f_{max} = 5512.5$Hz and 12 bins per octave. Finally, the chromagram is obtained by folding the constant-Q transformed spectrogram into one single octave to represent how energy is distributed among the 12 pitch classes. For 2), the chroma frames are aggregated with a median filter according to the beats found by a beat tracker [22] conforming to the music metrical grid. To have finer rhythmic resolution, each beat identified is spliced into two sub-beats before chroma frame aggregation. A final post processing step for 3) is applied to the sub-beat-synchronous chromagram by whitening it with a log function. The motivation of whitening is to boost the harmonic tones implied by the motives so that the difference between the same motive with and without harmonization is reduced. To consider transposition, the distance function used in *VMO* is replaced by a cost function having transposition invariance. To have a transposition invariant cost function, a cyclic permutation with offset k, on an n-dimensional vector $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})$ is defined as, $cp_k(\mathbf{x}) := \{x_i \rightarrow x_{(i+k \bmod n)}, \forall i \in (0, 1, \ldots, n-1)\}$, then the transposition invariant dissimilarity d between two vectors x and y is defined as, $d = \min_k\{\|x - cp_k(y)\|_2\}$. $n = 11$ for chroma vector and the cost function is used during the construction of *VMO*. A visualization of the result chromagram is depicted in the top plot of Fig 3.

3.2. Repeated Themes Discovery

For the specific task of repeated themes discovery, a pattern discovery algorithm is devised based on *VMO* and shown in Algorithm 1. The idea behind the algorithm is to track patterns by following `sfx` and `lrs`. `sfx` provides the locations of repeated suffixes and `lrs` contains the length for these repeated suffixes. In line 5 of Algorithm 1, state i is checked to make sure no redundant patterns are recognized and the lengths of patterns are larger than a user-defined minimum L. From line 6 to 10, the algorithm recognizes occurrences of established patterns and from line 11 to 15 it detects new patterns and stores them into $Pttr$ and $PttrLen$. Algorithm 1 returns $Pttr, PttrLen$ and K. $Pttr$ is a list of lists with each $Pttr[k], k \in \{1, 2, \ldots, K\}$, a list containing the ending indices of different occurrences of the kth pattern found. K is the total number of patterns found. $PttrLen$ has K values representing the length of the kth pattern in $Pttr$.

After the feature sequence O is extracted from the audio recording as described in the section 3.1, thresholds $\theta \in \{0.0, 0.001, 0.002, \ldots, 2.0\}$ are used to construct multiple *VMO*s with O, then the one *VMO* with the highest *IR* is fed into Algorithm 1 with L set to 5 empirically to find repeated themes and their occurrences. The result for finding repeated themes in one of the audio recordings from the dataset is shown in the bottom plot of Fig 3.

[1] Full experiment results and codes could be accessed at `https://github.com/wangsix/VMO_repeated_themes_discovery`

Algorithm	F_{est}	P_{est}	R_{est}	$F_{o(.5)}$	$P_{o(.5)}$	$R_{o(.5)}$	$F_{o(.75)}$	$P_{o(.75)}$	$R_{o(.75)}$	F_3	P_3	R_3	Time (s)
Proposed	**54.87**	68.93	55	**71.67**	77.34	67.09	**70.48**	72.7	68.55	**49.05**	62.59	49.77	**96**
[6]	49.8	54.96	51.73	38.73	34.98	45.17	31.79	37.58	27.61	32.01	35.12	35.28	454
[4]	23.94	14.9	60.9	56.87	62.9	51.9	–	–	–	–	–	–	–
[5]	41.43	40.83	46.43	23.18	26.6	20.94	24.87	32.08	21.24	28.23	30.43	31.92	196

Table 1. Results from various algorithms on the JKU Patterns Development Dataset. Scores are averaged across pieces.

Algorithm 1 Pattern Discovery using *VMO*

```
Require: constructed VMO, V, of length $T$ and a minimum pattern length $L$.
Ensure: $\texttt{sfx}, \texttt{rsfx}, \texttt{lrs} \in V$
Initialize $Pttr$ and $PttrLen$ as empty lists.
Initialize $prevSfx = -1, K = 0$
for $i = T : L$ do
    $pttrFound = False$
    if $i - \texttt{lrs}_V[i] + 1 > \texttt{sfx}_V[i] \wedge \texttt{sfx}_V[i] \neq 0 \wedge \texttt{lrs}_V[i] \geq L$ then
        if $\exists k \in \{1, \ldots, K\}, \texttt{sfx}[i] \in Pttr[k]$ then
            Append $i$ to $Pttr[k]$
            $PttrLen[k] \leftarrow \min(lrs[i], PttrLen[k])$
            $pttrFound = True$
        end if
        if $prevSfx - \texttt{sfx}[i] \neq 1 \wedge pttrFound == False$ then
            Append $\{\texttt{sfx}[i], i, \texttt{rsfx}[i]\}$ to $Pttr$
            Append $\min\{\texttt{lrs}[\{\texttt{sfx}[i], i, \texttt{rsfx}[i]\}]\}$ to $PttrLen$
            $K \leftarrow K + 1$
        end if
        $prevSfx \leftarrow \texttt{sfx}[i]$
    else
        $prevSfx \leftarrow -1$
    end if
end for
return $Pttr, PttrLen, K$
```

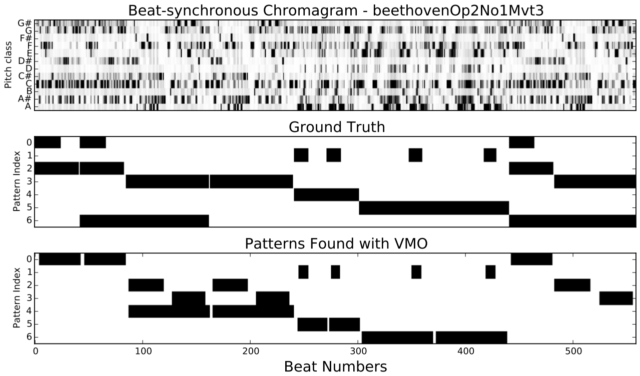

Fig. 3. (Top) Beat-synchronous Chromagram, (Middle) Ground truth from JKU dataset. (Bottom) Found patterns by Algorithm 1.

4. EVALUATION

The evaluation follows the metrics proposed in the Music Information Retrieval Evaluation eXchange (MIREX) [1]. Three metrics are considered for inexact pattern discovery. For each metric, standard F_1 accuracy score, defined as $F_1 = \frac{2PR}{(P+R)}$, precision P and recall R are calculated. The first metric is the establishment score (est) which shows how each ground truth pattern is identified and covered (taking inexactness into account, not considering occurrences) by the algorithm. The second metric is the occurrence score ($o(c)$) with a threshold c. The occurrence score measures how well the algorithm performs in finding occurrences of each pattern. The threshold c determines if an occurrence should be counted or not. The higher the c, the lower the tolerance. $c = \{0.5, 0.75\}$ are used in standard MIREX evaluation. The last metric is the three-layer score that considers both the establishment and occurrence score. The results of the proposed framework are listed in Table 1 with comparison to previous works.

From Table 1, it is clear that the proposed framework significantly improves the state of art reported in [6] in all metrics. The significant improvements of the F_1 scores indicates that the proposed framework is capable of finding both correct patterns and their occurrences without biasing from either type-1 or type-2 errors. In addition to the improvement of accuracy, the proposed framework is also significantly faster than other approaches as reported in Table 1. The computation is timed the same way as reported in [6], which excludes the calculation of sub-beat-synchronous chroma and only measures on the pattern discovery routines. In this paper, it includes finding the optimal θ value, construction of *VMO* and Algorithm 1. One major reason causing less computation time is that the proposed framework avoids the calculation of self-similarity matrix [6, 21] of the chromagram.

5. CONCLUSIONS

In this paper, a framework for discovering patterns embedded in a signal is proposed and shown to improve the state of the art significantly for the task of discovering repeated themes from audio recordings. The core of the framework is *VMO*, a data structure equipped with model selection criterion derived from music information dynamics and capable of symbolizing a signal while keeping its temporal dynamics. From the experiment result of this paper, it is shown that *IR* is indeed a good indication for temporal structuredness. The combination of *IR* and *VMO* is shown to have potentials in unsupervised structural discovery problems as well.

6. REFERENCES

[1] Tom Collins, "Discovery of repeated themes and sections," *Retrieved 4th May, http://www.music-ir.org/mirex/wiki/2013:Discovery_of_Repeated_Themes_&_Sections*, 2013.

[2] B Janssen, WB Haas, A Volk, and P Kranenburg, "Discovering repeated patterns in music: potentials, challenges, open questions," in *10th International Symposium on Computer Music Multidisciplinary Research*. Laboratoire de Mécanique et d'Acoustique, 2013.

[3] 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.

[4] Tom Collins, Sebastian Böck, Florian Krebs, and Gerhard Widmer, "Bridging the audio-symbolic gap: The discovery of repeated note content directly from polyphonic music audio," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[5] Oriol Nieto and Morwaread Farbood, "MIREX 2013: Discovering musical patterns using audio structural segmentation techniques," *Music Information Retrieval Evaluation eXchange, Curitiba, Brazil*, 2013.

[6] Oriol Nieto and Morwaread Farbood, "Identifying polyphonic patterns from audio recordings using music segmentation techniques," in *The 15th International Society for Music Information Retrieval Conference*, 2014.

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

[8] Shlomo Dubnov, "Unified view of prediction and repetition structure in audio signals with application to interest point detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 327–337, 2008.

[9] 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.

[10] Marcus T Pearce and Geraint A Wiggins, "Auditory expectation: The information dynamics of music perception and cognition," *Topics in cognitive science*, vol. 4, no. 4, pp. 625–652, 2012.

[11] Keith Potter, Geraint A Wiggins, and Marcus T Pearce, "Towards greater objectivity in music theory: Information-dynamic analysis of minimalist music," *Musicae Scientiae*, vol. 11, no. 2, pp. 295–324, 2007.

[12] Marcus T Pearce, Daniel Müllensiefen, and Geraint A Wiggins, "Melodic grouping in music information retrieval: New methods and applications," in *Advances in music information retrieval*, pp. 364–388. Springer, 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] Arnaud Lefebvre, Thierry Lecroq, and Joël Alexandre, "An improved algorithm for finding longest repeats with a modified factor oracle," *Journal of Automata, Languages and Combinatorics*, vol. 8, no. 4, pp. 647–657, 2003.

[15] Shlomo Dubnov, Gerard Assayag, Arshia Cont, et al., "Audio oracle: A new algorithm for fast learning of audio structures," in *International Computer Music Conference*, 2007.

[16] Arshia Cont, Shlomo Dubnov, Gérard Assayag, et al., "Guidage: A fast audio query guided assemblage," in *International Computer Music Conference*, 2007.

[17] Greg Surges and Shlomo Dubnov, "Feature selection and composition using pyoracle," in *The 9th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2013.

[18] Cheng-i Wang and Shlomo Dubnov, "Guided music synthesis with variable markov oracle," in *The 3rd International Workshop on Musical Metacreation, 10th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

[19] Cheng-i Wang and Shlomo Dubnov, "Variable markov oracle: A novel sequential data points clustering algorithm with application to 3d gesture query-matching," in *International Symposium on Multimedia*. IEEE, 2014, pp. 215–222.

[20] Arnaud Lefebvre and Thierry Lecroq, "Compror: online lossless data compression with a factor oracle," *Information Processing Letters*, vol. 83, no. 1, pp. 1–6, 2002.

[21] Juan Pablo Bello, "Measuring structural similarity in music," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2013–2025, 2011.

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