# Context-Aware Hidden Markov Models of Jazz Music with Variable Markov Oracle

**Cheng-i Wang, Shlomo Dubnov**
University of California, San Diego

**Abstract**

In this paper, a latent variable model based on the Variable Markov Oracle (VMO) is proposed to capture long-term temporal relationships between sequential observations. The proposed latent variable model is a multi-level Hidden Markov Model (HMM) extracted from the VMO, and is called the VMO-HMM. The musical importance of the proposed model is demonstrated by modeling harmonic progressions of jazz music. The VMO-HMM is able to reveal functional harmony relations beyond chord labels, provides theoretic insight into scale-chord relations and aids jazz harmony improvisation as a model of music meta-creation.

## 1 Introduction

No mater if it is a musicologist analyzing a jazz piece for theoretical understanding or a musician comprehending a jazz lead sheet for performances, it is undoubtedly that there is always more to interpret than just following the chord label sequence given on the lead sheet. For example, a musician has to decide on the harmonic function of a chord given its neighboring chords, then gathers hints from extended chord labels and chooses an appropriate yet interesting scale to perform over that chord. In that sense, the same chord labels in a tune could serve different harmonic functions given different surrounding chords and lead to different note choice possibilities.

Most of the computational models using harmonic content of music pieces view chord labels as the latent variables (Sheh and Ellis 2003; Eigenfeldt and Pasquier 2010) emitting surface musical events such as notes. Among various latent variable models, HMM (Allan and Williams 2004; Nakamura et al. 2015) is used extensively since it could be used to model the sequential relationships between successive chord labels. But in those models, each chord label is modeled by one latent variable, thus it is difficult for those models to distinguish among different harmonic functions of the same chord labels. In (Gillick, Tang, and Keller 2010), markov chains are learned to enhance the modeling of jazz melodies for style consistent music generation but harmonic structures are not modeled.

Motivated by the above observation, a latent variable model is proposed to model the relationships between observations and latent variables, but also to split latent variables with different contexts having similar observations. Furthermore, from a musical standpoint, the interpretive nature of a score description (such as chord labels in a lead sheet) and its relationship to the realized music could be described by the chord-scale theory and realized by the proposed latent variable model.

### 1.1 Musical Theory of chord-scale improvisation

The problem of assigning pitch collections to chord symbols is a fundamental problem in creative or improvisatory interpretation of Jazz standards, figured bass, tablature notations and most of the popular music chord notations. Such partial chord specifications provide, to a different level of specificity, an overall sense of harmonic structure that the music composition should follow, leaving great amount of liberty to the improviser or arranger to decide on the actual realization of notes, voices and rhythms. In many respects, composing a chord or harmonic label sequence is a meta-composition process that has to be further completed during the actual performance by musicians.

One musical theory that tries to address the question of note choices given chord labels is the so called chord-scale theory (Nettles and Graf 1997). A scale is a series of pitches ordered by relative height or frequency, often limited to a subset from the complete set of twelve chromatic notes. An important aspect of a scale is its ordering of notes that induces a sense of stepwise proximity that is different from chromatic semitone ordering. Such pitch organization introduces a sense of correct and false notes, or notes that should be avoided when two or more tones are played simultaneously. As the harmonic complexity grew with the advent of chromaticism in classical and jazz music of the 20th century, composers approached this chromatic freedom in new ways that challenged the notion of harmony and traditional scales, such as the twelve-tone and serialist techniques. Other composers

preserved more traditional relations between chords and scales while significantly expanding the choices of chords, and thus also constructing new and more complicated scales.

This evolution of musical language presents new opportunities in musical material constructions, with scales mediating between the musical surface events[1] and the underlying harmonic structure. Such a scale-oriented musical practice allows a more hierarchically structured compositions where contextual relations may exist both on the surface and on a latent harmonic level. Due to the large variety and originality of establishing chord-scale relations by different composers, this musical problem of describing the relations between surface and structure has been difficult to formalize. Some theories try to explain the dynamics of moving between scales based on shared subsets or efficient voice leading (Tymoczko 2004). For example, the C diatonic scale can be linked to G diatonic scale by "maximally smooth voice leading" since the two scales share six common tones and moving from one scale to another only requires a shift of a single pitch class by a semitone.

The interesting point in considering chord relations through scale changes and voice leading is that traditional harmonic relations, such as dominant-tonic or other functional harmony considerations, could be revisited, generalized, and even contradicted by following scale dynamics. In VMO-HMM, a somewhat similar construction occurs by clustering notes using temporal consideration, with latent states serving the role of a mediating structure between chord labels and the musical surface. This allows richer possibilities for rendering musical surface by alternative choices from latent states that are richer than ordinary chord labels. The system allows construction of altered chord progressions and improvisation from a vocabulary of note selections (scales) by recopying relevant musical excerpts from a database. These applications of random chord sequence generation and improvisation for a given chord progressions (query based improvisation) will be demonstrated in the paper.

### 1.2 Context-Aware Hidden Markov Models

A Variable Markov Oracle (VMO)(Wang and Dubnov 2015; Wang, Hsu, and Dubnov 2016) is a data structure based on the Factor Oracle automaton that is capable of identifying repeated subsequences within a multivariate time series and was used for machine improvisation and multimedia analysis. It is shown in (Wang and Dubnov 2015) that a VMO is capable of clustering data points of a multivariate time series based on their temporal relations, and tracking the sequential transitions between these clusters. This effectively makes VMO into a latent temporal model with special behavior that is different from the common HMM due to its variable length modeling property. In this paper, a further study of how a VMO differs from a HMM is presented. A visual demonstration of the context aware aspect of VMO-HMM is depicted in figure 1. In the example these observations are clustered into groups according to different clustering methods. Figure 1 shows these results by marking the cluster assignment with different colors. Examples of time series modeled by VMO, HMM Gaussian Mixture Models (HMM-GMM) and K-Means are shown. It should be noted that the K-Means clusters observations based on spatial positions only (in the feature space), and that both VMO and HMM-GMM take the time trajectories into account. In these examples it is evident that only VMO is capable of distinguishing between observations that are spatially but not sequentially (temporally) close and assigning them to different clusters. After examining the results it is clear that the establishment of clusters by HMM-GMM and K-means is mainly determined by spatial relationships between observations but not temporal relationships. Although the possibility of forming a latent model was discussed in previous studies of VMO and Viterbi-like algorithms were proposed for different applications, the VMO data structure still has to be kept in order for the algorithms to work. In this paper, the proposed latent model is a compact version of the VMO data structure with a novel statistical model and probabilistic interpretations.

In section 2, the VMO data structure and the method of extracting a latent model using VMO are introduced. Jazz music analysis is described in section 3. Novel music creation possibilities using VMO-HMM is proposed in section 4. Conclusion and discussions are elaborated in section 5.

## 2 Variable Markov Oracle as Latent Model

VMO was introduced in (Wang and Dubnov 2014) as an extension to the Audio Oracle (AO) (Dubnov, Assayag, and Cont 2007) having the capability of guided machine improvisation. A VMO could be viewed as an on-line clustering algorithm without the need to specify the number of clusters. A VMO could also be considered as a data structure that traces repeated sub-sequences in the latent space. These two properties make VMO capable of both modeling and generating multivariate time series. The technical details of constructing a VMO could be found in (Wang, Hsu, and Dubnov 2016) and are not repeated here.

The clustering property of a VMO is explained in detail in (Wang, Hsu, and Dubnov 2016). The VMO was introduced as a data structure that allowed symbolization and clustering, but without an underlying statistical model. In (Wang, Hsu, and Dubnov 2016) an HMM analogy was made as the first attempt to establish a statistical model for VMO. It took into account the inference of emission probabilities from observation and but did not model transition probabilities. The current work is the most complete analogy or statistical interpretation of the FO structure after IR optimization to an HMM.

In short, a VMO records where and how long the longest repeated suffixes happened for every time step in the time

[1] By musical surface events we mean a full specification of sound events, represented as information that is complete enough to allow a performable musical score.

series, and stores them in two arrays, `sfx` and `lrs` respectively. Since for each observation only one longest repeated suffix is recorded, each observation in a time series is indexed by VMO by assigning a label that is based on the unique paths that are defined by suffix records. The labels are stored in an array called `latent`. The suffix records are called *suffix links* since they indicate connections between observations if they share similar context (suffixes). Observations assigned the same label possess two properties that are utilized in this paper: The first one is that the distances between the observations connected by suffix links are below a found threshold $\theta$ during the model selection process. The second one is that they all share common suffixes in the latent space. The first property sets the basis of modeling observations with a latent variable, the second property provides a Markovian relationships between the latent variables.

## 2.1 VMO-HMM

The HMM-like model extracted from VMO is called the VMO-HMM. To extract a VMO-HMM from a VMO, each latent variable is represented by the centroid extracted from the clustered observations. The choice of centroid could be flexible, such as mean or median, depending on the applications. To extract the Markov transition probabilities from a VMO, the `lrs` array is used. The `lrs` arrays contains the

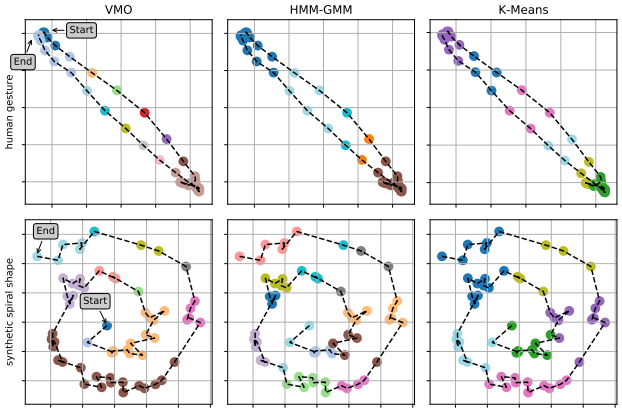


Figure 1: Two time series examples modeled by three different approaches. The gesture in the top row is a real world 3-D human skeletal joints gesture projected onto its first two principle components. The bottom one is a synthetic spiral sequence. The bottom one is a shape commonly tested in manifold discovering and clustering. Three approaches are used to model these time series. (From right to left) K-Means, HMM-GMM and *VMO*. Each observation along the time series is represented as a colored circle with its color represents the label (hidden/latent state) that the observation is assigned to. observations with the same color in the same plot belong to the same label. The starting and ending positions of the two time series are annotated in the left most column. Dashed lines connecting the data points represent the time progression trajectories of each time series.

lengths of the longest repeated suffixes, thus it also provides variable-length Markov transition information. To obtain these information, a 3-D Markov transition tensor is created instead of a 2-D Markov transition matrix. In algorithm 1, a simple algorithm is provided here to show how the tensor is extracted.

In algorithm 1, the counts of occurrences between consecutive latent variables are accumulated across the first dimension of $S$, with the index if the first dimensiton representing the order of each Markov transition matrix. In following sections, musical analysis and creation utilizing both the latent variables and the Markov tensors extracted from VMO will be shown.

## 2.2 Model Selection

Constructing a VMO with different distance threshold $\theta$ values will result in VMOs with different suffix and latent variable structures. To select the one latent variable model with the most informative variable-order Markov structure, Information Rate (IR) is used as the criterion in model selection between different structures generated by different $\theta$ values. IR is an information theoretic measure capable of measuring the information content of a time series (Dubnov 2006). 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 latent variable sequence.

The IR calculation for constructing a VMO is the same as that for an AO (Dubnov, Assayag, and Cont 2011). Let $x_1^N = \{x_1, x_2, \ldots, x_N\}$ denote time series $x$ with $N$ observations and $H(x)$ the entropy of $x$. The definition of IR is

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

The value of IR could be approximated by replacing the entropy terms in (1) with a complexity measure, $C(x)$, associated with a compression algorithm. This complexity mea-

**Algorithm 1** HMM tensor extraction

**Require:** An indexed VMO $V$, max variable length $M$
$N \leftarrow$ the number of latent variables in $V$
Create a 3-D tensor $S$ with dimensions $\{M, N, N\}$
$T \leftarrow$ the number of data points in $V$
**for** $t = 2 : T$ **do**
  $i \leftarrow \texttt{latent}_V[t-1]$
  $j \leftarrow \texttt{latent}_V[t]$
  **if** $\texttt{lrs}_V[t] < 2$ **then**
    $S[1, i, j] \mathrel{+}= 1$
  **else**
    $S[1 : \texttt{lrs}_V[t] - 1, i, j] \mathrel{+}= 1$
  **end if**
**end for**
**for** $m = 1 : M$ **do**
  Normalize each row in $S[m, :, :]$
**end for**
**return** $S$

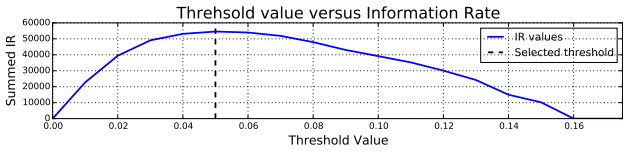


Figure 2: IR values are shown on the vertical axis while $\theta$ are on the horizontal axis. The solid blue curve shows the relationship between IR and $\theta$, and the dashed black line indicates the chosen $\theta$ by locating the maximum IR value. Intuitively, $C(x_n)$ and $C(x_n|x_1^{n-1}$ exhibit an inverse relationship with increasing $\theta$, which leads to the IR curve showing the quasi-concave function shape, and allows a global maximum be located.

sure is the number of bits used to compress $x_n$ independently using the past observations $x_1^{n-1}$.

$$IR(x_1^{n-1}, x_n) \approx C(x_n) - C(x_n|x_1^{n-1}). \quad (2)$$

Compror, a lossless compression algorithm based on the FO and the lengths of the longest repeated suffixes (`lrs`), is provided in (Lefebvre and Lecroq 2002). The detailed formulation of combining Compror, the AO structure and IR is provided in (Dubnov, Assayag, and Cont 2011; Wang and Dubnov 2015). Basically, a reasonable range of $\theta$ values are used to create multiple VMOs indexing the target signal, then the one VMO that returns the highest IR value will be the one VMO being used. The found $\theta$ value determines if an observation belongs to a cluster or not, while the actual label assigned to a cluster is determined by the context (repeated suffixes) in the latent variable space. In other words, observations that are spatially close could be assigned to different clusters if their preceding observations (in time) came from different contexts (repeated suffixes). A visualization of the sum of IR values versus different $\theta$s is depicted in figure. 2.

## 3 Music Analysis

To exemplify the use of VMO-HMM on musical applications, a case study on analyzing Jazz music harmonic progression is conducted. The piece being analyzed is "Now's the time" from Charlie Parker. The lead sheet in MusicXML and an accompaniment recording in midi are both available. To analyze the harmonic progression automatically, the MIDI accompaniment is used as the input to VMO. The lead sheet comes with the melody and human annotated chord labels, and serves as the reference to the VMO analysis.

To obtain harmonic information from MIDI, the MIDI note events are first quantized to a piano roll matrix with dimension $\{128, B\}$, where $B$ stands for the number of bars and the first dimension for MIDI pitch values. The values in the piano roll matrix are velocities ranging between $[0, 127]$ with 0 representing a none event. There are several choices for the note velocity aggregation (pooling) within a bar, such as max, mean or median. For this case study, max-pooling is used to aggregate the note velocities within each bar along the time axis. Since the time signature and tempo are given in the MIDI file, the bar locations could easily be determined. After extracting the piano roll matrix, it is further folded over octaves to form a chroma-like matrix (midi-chromagram) with dimensions $\{12, B\}$, with the first dimension represents the pitch class {C,Db,D,Eb,E,F,Gb,G,Ab,A,Bb,B}. A normalization along the time axis is done to normalize the values of each pitch class to be between $[0, 1]$. A visualization of the final midi-chromagram matrix is depicted in Figure 3.

To further strengthen the harmonic modeling, the 12-dimension data points from the midi-chromagram are projected onto the tonnetz space (Morris 1998) during the construction of a VMO with the midi-chromagram. The $L-2$ norm distance is used during the construction. After indexing the midi-chromagram with a VMO, the clusters could be retrieved by grouping the frames in `latent` having the same label. The clusters of chroma frames for each clusters formed by VMO could be visualized as in Figure 4.

Parts of the reference lead sheet with chord labels are shown in figure 5. To examine how well a VMO-HMM capture the relationships between observations and latent variables. A qualitative comparison between the reference chord labels and the clusters from the VMO-HMM shows that the discovered clusters from the VMO-HMM do capture harmonic meanings and provide more information than the reference chord labels from the lead sheet. Comparing the centroid chroma (right column) in figure 4 with the score in figure 5, cluster-0 and cluster-5 match with chord labels `F` and `F7`, cluster-1 matches with `B♭7`, cluster-2 with `Am/D7`, cluster-3 with `Gm`, cluster-4 with `C7` and cluster-6 with `C7(♭9)`. Since cluster-7 only has only 1 frame in it and paragraph considerations, its discussion will be skipped. The unsupervised discovery with the VMO-HMM is nearly perfect to the reference chord labels, which shows that the VMO-HMM is capable of capturing the groupings of midi-chroma frames into harmonically meaningful clusters. Furthermore, an interesting split of the chord label `[F,F7]` into cluster-0 and cluster-5 shows that the VMO-HMM does capture a level deeper than what the surface chord labels suggest. By examining where cluster-0 and cluster-5 are located in the score, the `F7`s associated with cluster-0 are undeniably the tonic with a flat 7 for which Major or Mixolydian scales could be suitable. On the other hand, the `F7` associated with cluster-5 are passing chords between

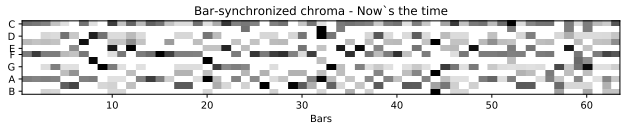


Figure 3: The midi-chromagram extracted from the MIDI accompaniment recording of the Jazz piece "Now's the time".

C7. Like the one in measure 11. Though it is marked as F7 in the lead sheet, in the accompaniment MIDI file it is actually played as F13(♯11), in which a Lydian Dominant scale could be used.

Finally, the Markov transition matrices obtained from algorithm 1 of different orders are shown in figure 6. The $j$th entry in the $i$th row in each matrix represents the probability from cluster-$i$ transition to cluster-$j$. Matching the clusters to the chord labels verifies that this piece follows a tight [ii,V,I] progression of the chord sequence [D7, Gm, C7] and [Gm, C7, F7]. By examining these Markov transition matrices, it could be observed that the higher the order, the sparser the Markov transition matrix is. As the transition matrix gets sparser, the more definite the transition probabilities and fewer choices for possible next states. The musical interpretation combining the observations of the transition matrices and the clusters is that lower order Markov transition matrices capture freer and more jazzy relationships between chord clusters while higher order ones capture stricter harmonic tonal function relationships between chord clusters. For example, in the 1st-order transition matrix, cluster-0, which matches to [F,F7], could transition to any of the other clusters except for cluster-5, whose chroma content is similar to cluster-0. But as the order gets higher, the possible clusters that cluster-0 could transition to becomes fewer, and converges to cluster-1 and cluster-2, which matches to B♭7 and Am/D7 respectively. The converged transitions between F7 to B♭7 and Am/D7 are standard [I,IV] or [I, vi] movements in tonal theory.

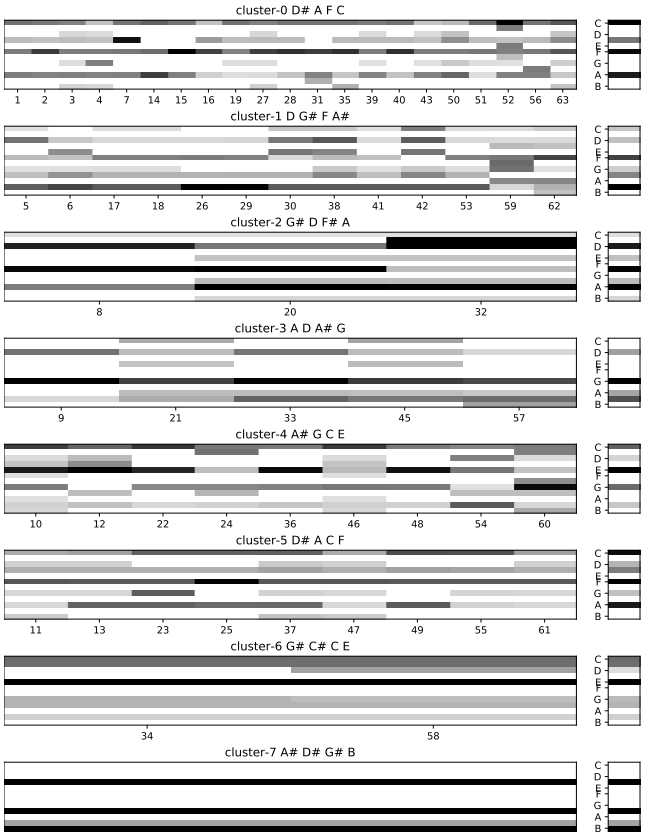


Figure 4: The clusters formed by the VMO on "Now's the time". The left matrix of each row is the collection of frames from the midi-chromagram, the right single vector is the median centroid obtained by median-pooling along the time axis within that cluster. The four pitches for each cluster represent the four most dominant pitches within that cluster by velocity.

## 4 Music Improvisation

The use of a VMO for improvisation and synthesis is already introduced in (Wang, Hsu, and Dubnov 2016). Previous works focused on guided improvisation and synthesis using a VMO directly on audio signals. The guided music generation was made possible by specifying a query to recombine the indexed audio signal based on the VMO suffix structures. The limitation for previous works is that the query and the target (VMO-indexed) signals have to use the same alphabet, or in other words, the same feature or type of signal. In this paper, a framework analyzing symbolic music representation is proposed in section 3, thus allowed the VMO to further expand its generative aspect to symbolic representations. The most important advancement of using symbolic representation with a VMO is that it allows the user to specify a query signal that uses a different alphabet from the target signal. To be more specific, the user could now specify a chord label sequence as input to the improvisation system. The system then translates the chord label se-

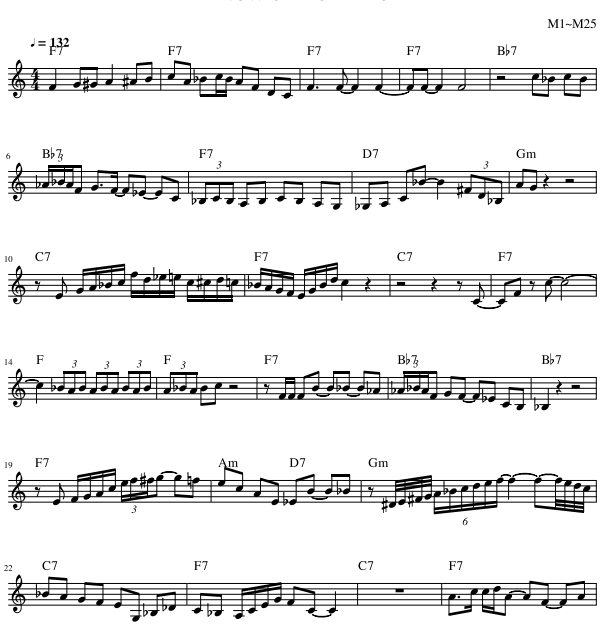


Figure 5: Partial score from the lead sheet of "Now's the time" from bar 1 to 25.

quence to pitch class profiles, which is essentially the same representation as the midi-chromagram. Then the translated sequence is used as the query to the VMO to generate a new sequence in the same manner as proposed in (Wang, Hsu, and Dubnov 2016).

### 4.1 Jazz chord sequences

Musicologists have tried to capture the essence of jazz chord sequences by applying rules of classical harmony to understand how basic harmonic structures have been transformed in a jazz composition. One common technique of chord substitution rule can be formalized as a "rewriting rule" which allows transforming a subsequence of chords into another subsequence of chords that introduces diversity, without, in principle, changing the harmonic function of the subsequence. Indeed, although jazz harmony could be considered born from Classical harmony in an evolutionary viewpoint, the harmonic functions of jazz chords seem to be much more complex than those in Classical four-part chorals, because of the underlying combinatorial "game" at play. For example, in classical theory the chord of `C` major and `F#` chord are the most "distant" in terms of their tonal context. In jazz however, a `C(7)` and `F#(7)` are closely related through sharing a common tri-tone axis, and may be considered interchangeable. Another common example of a distinction between classical and jazz interpretation of chords is the functional role of `C` and `C7`. While in classical harmony a `C7` is considered an unstable dominant chord that is expected to resolve to an `F`, in jazz, `C` and `C7` are often considered equivalent. These examples indicate that the harmonic rules which make sense in classical harmony might not be strictly obeyed in other tonal or modal musical styles. In a VMO-HMM model, the relations between harmonic constructs, captured by the latent variables, depend on the previous note aggregation phase (the feature extraction part described in sections 3) that is based on surface level note dynamics. Our experiments show that there are two main transition types between

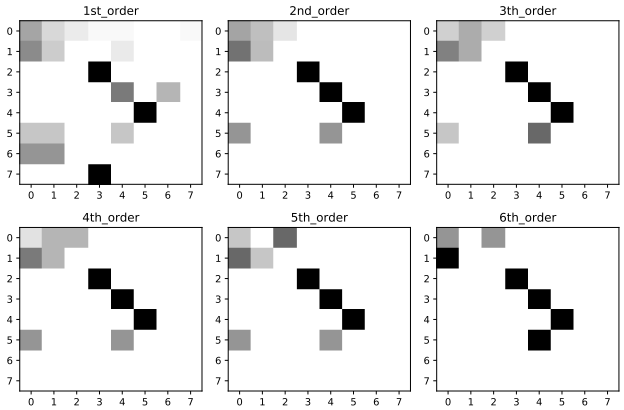


Figure 6: Different Markov transition matrices extracted from the VMO-HMM, from lower to higher order. The entry $[i, j]$ in each matrix represents the probability from latent variable $i$ to $j$.

latent states suggesting different tonal relations and chord transitions. One of them is the common jazz operation called the "enrichment" of chords, either viewed as 7, 9, 11 and 13 notes, or as sustained or color notes. These enrichments are often used as special events, and understanding the context of their application is important for allowing targeted and effective use of such harmonic devices. In the analysis conducted in section 3 it is found that the same musical notations (such as the `F7` chord in "Now's the time" example) could be split between two clusters, where one (cluster-5) of them contains a particularly more rich and embellished set of notes than the other. Accordingly, when a VMO-HMM is used to generate a new chord progression by a random walk on the Markov structure between the latent variables, such alterations, substitutions or enrichments may be controlled as part of the musical meta-composition design.

### 4.2 Random Walks on VMO-HMM

Given different Markov transition matrices from different `lrs` values, it is straight forward to sample latent variable sequences treating each row in the transition matrix as a multinomial distribution conditioned on its previous latent variable. Continuing from the analysis example used in section 3, given cluster-0 as the fixed initial latent variable, each next latent variable is drawn randomly given the multinomial distribution conditioned on the current latent variable. After the latent variable sequences are sampled, chord labels are assigned to latent variables based on the clusters shown in figure 4 in the same way as section 3. By examining the two sampled examples (figure 7), it could be observed that the chord choices and temporal relations of the lower order one are freer than the higher order one. If we focus only on the root progression of these two example sequences, the 1st-order sequence contains progressions such as `[I,ii,VI]`, `[I,vi,ii,V]` and `[IV,V,I]`, while the 5th-order one contains mainly the `[I,vi,ii,V,I,V]` progression. Based on these observations, the lower order Markov model indeed captures a wider variety than the higher order Markov model but lacks repetitive structures. On the other hand, the higher order Markov model captures more salient harmonic progressions in the music spanning multiple bars. To render actual musical content from the chord label sequences, one simple method is to randomly select a bar containing the midi events from the cluster associated to the latent variable. Due to space limitation the generated scores are not shown here and could be found in the repository[2]. It should be noted that since the random sampling is on the latent variable space, there is no one fixed answer for the musical realization given the sampled latent variable sequence. Reshuffling the original musical content is just a convenient way, other generative methods based on chord labels could also be used.

### 4.3 Query VMO-HMM by Chord Label Sequence

As mentioned in the beginning of this section, the other advantage of using the VMO-HMM is that it provides a

[2] `https://github.com/wangsix/markov_improvisation`

| F7 | F7 | F7 | Gm |C7(♭9)| F7 | D7 | Gm |C7(♭9)| Bb7 | C7 |F7(♯11)| C7 |

(a) 1st-order

| F7 | D7 | Gm | C7 |F7(♯11)| C7 |F7(♯11)| F7 | D7 | Gm | C7 |F7(♯11)| C7 |

(b) 5th-order

Figure 7: Two sampled 12-bar chord label sequences of different orders Markov transition matrices from the VMO-HMM on the piece "Now's the time". It should be noted that the chord labels are inferred by human inspection on the clusters shown in figure 4, not the chord labels from figure 5.

complete setting for ordinary HMM Viterbi recognition algorithm (Sheh and Ellis 2003). A Viterbi algorithm using VMO was proposed in (Wang and Dubnov 2015), where the transition probabilities are assumed to be a uniform distribution on the forward links from a frame to possible next frames. In the VMO-HMM setting, the transition probabilities between latent variables are learned from the oracle structure based on the longest repeated suffixes of each frame.

To use the Viterbi algorithm with a VMO-HMM for music content generation, one can specify a chord label sequence similar to the generated sequences in figure 7, then translate the chord label sequence into a chroma vector sequence. To translate chord labels to chroma vectors, one can simply use 12-dimensional binary vectors to represent 12-tone-pitch classes. An example of such translation and comparison to the actual chroma sequence is shown in figure 8. The translated chroma sequence is then used as the observation in the Viterbi algorithm. To infer the latent variable sequence generating the observations, the emission probability of an observation generated by a latent variable could also be simplified as the cosine distance between the binary pitch class vector (observation) and the centroid (mean or median of the cluster) chroma vector normalized to be a positive-valued vector which sums to 1. The Viterbi algorithm decodes an observation sequence to a latent variable sequence. The decoded latent variable sequence could then be used to generate new musical materials as in section 4.2. As a proof of concept using the aforementioned approach generating musical contents with a user specified chord label sequence, the chord labels from the first 12 bars in the reference MusicXml file of "Now's the time" is used as the input chord label sequence to the Viterbi algorithm to see if the decoded latent variable sequence matches the given chord labels. The goal of this proof of concept is to see if given a version of translation between the chord labels and the pitch class profiles, how well could the Viterbi decoder from a VMO-HMM work. The result of this initial attempt works well since the Viterbi decoder is capable of finding the exact latent variable sequence as the input chord label given the reduced representation from a VMO-HMM. The testing script can also be found in the repository provided above. In figure 9, both the query and the decoded chord label sequences are shown. At bar 11, although the input label from the lead sheet specifies `F7`, but the Viterbi algorithm with the VMO-HMM extracted from the MIDI accompaniment file is able to spell out `F7(♯11)` given its different context from earlier $F7$s. This observation confirms that the VMO-HMM is capable of distinguishing similar chord given their pitch classes if they have different context in the music.

## 5 Discussion and Conclusion

In this paper we presented a method of clustering collections of notes according to similarities in their temporal context, and learning multiple transition probabilities between the clusters arranged into a Markov tensor indexed by the length of longest repeated sequences. This latent construction with variable memory property is called the VMO-HMM. A musical theoretic interpretation of the latent states was derived by observing the relations between cluster contents and chord labels for a piece of jazz tune. It is suggested that the latent states could be considered as a generalization of the scale-chord theory where the same chord labels could be "split" into multiple possible choices of scale renderings. Moreover, finding the optimal transition path between latent states creates alternative harmonizations for a chord sequence inputting into the VMO-HMM. This effectively creates an enrichment of the harmonic language that is learned idiomatically from musical MIDI recordings.

One other interesting interpretation of the proposed model comes from a music cognition standpoint. A particularly intriguing aspect of the VMO-HMM Markov tensor representation is that it explicitly models the relations between Markov statistics and the lengths of musical memories that were used to learn these statistics. The tensor representation makes evident the dependency between memory length (`lrs`) and the possibilities it offers for continuation, which seems to suggest a correspondence between anticipations and the type of memory involved in musical practice. It was suggested long ago by Meyer (Meyer 1956) that "the same physical stimulus may call forth different tendencies in different stylistic contexts ... For example, a modal cadential progression will arouse one set of expectations in the musical style of the sixteenth century and quite another in the style of the nineteenth century.". Music cognition researchers (Huron 2006) take this approach a step further by suggesting distinct cognitive mechanisms, possibly even brain ones, for different types of

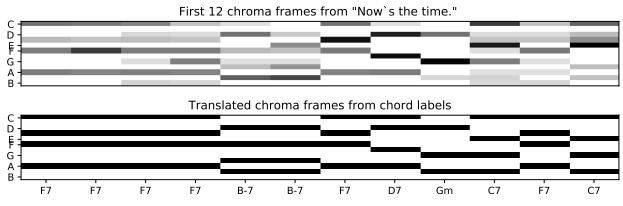


Figure 8: The actual midi-chromagram from the MIDI accompaniment compared to the translated query chromagram (pitch classes) from chord labels in the lead sheet.

| F7 | F7 | F7 | F7 | B♭7 | B♭7 | F7 | D7 | Gm | C7 | F7 | C7 |

(a) Query Chord Labels

| F7 | F7 | F7 | F7 | B♭7 | B♭7 | F7 | D7 | Gm | C7 | F7(♯11) | C7 |

(b) Decoded Chord Labels

Figure 9: The query and decoded chord label sequences. All the chord labels are matched besides bar 11, where the `F7` from the lead sheet is identified as `F7(♯11)` given the VMO-HMM.

musical memory responsible for veridical versus schematic expectations. Veridical expectation is an expectation that arises due to knowledge about a specific stimulus, such as memorizing a musical piece. Schematic expectation arises from general mental schema related to a certain style of music. So the reason that a deceptive cadence `[V-vi]` evokes a physiological response characteristic of surprise even when the listener is familiar with a piece, is that the fast (schematic) brain is being surprised by the "deception" while the slow (veridical) brain is not. One hypothesis is that separate mechanisms are required for learning veridical expectations that require greater familiarity with the piece, and schematic expectations that are much shorter and have more possible continuations. We suggest that the use of `lrs` dimension as a control parameter for the level of musical confidence is distinct from surprisal measures that are based on counts of future branching choices. Navigating between more certain veridical memory paths versus less probable sequences may reflect on differences in mental flexibility and confidence of making musical choices, and could be a valuable meta-creation parameter motivated by theories of musical cognition.

In summary, the oracle structure has been used extensively for machine improvisations. The VMO-HMM provides a more compact and abstract representation of the oracle structure while keeping its variable-length Markov properties. The tensor representation also presents opportunities for consolidating different VMOs from different music pieces into one unified model for a corpus. A few things have to be taken care of for such unification, such as normalizing the key for each song so that the functional harmonic relationships between pitch classes are consistent and matching clusters of chroma between different songs. Once these steps are dealt with, a system that takes chord label sequences as input and output musical content that takes both veridical and schematic aspects of music anticipations into account using the VMO as the core engine could be devised.

## Acknowledgment

This research is supported by CREL research grant from the UCSD Office of Graduate Research.

## References

[Allan and Williams 2004] Allan, M., and Williams, C. K. 2004. Harmonising chorales by probabilistic inference. In *NIPS*, 25–32.

[Dubnov, Assayag, and Cont 2007] Dubnov, S.; Assayag, G.; and Cont, A. 2007. Audio oracle: A new algorithm for fast learning of audio structures. In *Proceedings of International Computer Music Conference (ICMC)*.

[Dubnov, Assayag, and Cont 2011] Dubnov, S.; Assayag, G.; and Cont, A. 2011. Audio oracle analysis of musical information rate. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, 567–571. IEEE.

[Dubnov 2006] Dubnov, S. 2006. Spectral anticipations. *Computer Music Journal* 30(2):63–83.

[Eigenfeldt and Pasquier 2010] Eigenfeldt, A., and Pasquier, P. 2010. Realtime generation of harmonic progressions using controlled markov selection. In *Proceedings of ICCC-X-Computational Creativity Conference*, 16–25.

[Gillick, Tang, and Keller 2010] Gillick, J.; Tang, K.; and Keller, R. M. 2010. Machine learning of jazz grammars. *Computer Music Journal* 34(3):56–66.

[Huron 2006] Huron, D. B. 2006. *Sweet anticipation: Music and the psychology of expectation*. MIT press.

[Lefebvre and Lecroq 2002] Lefebvre, A., and Lecroq, T. 2002. Compror: on-line lossless data compression with a factor oracle. *Information Processing Letters* 83(1):1–6.

[Meyer 1956] Meyer, L. B. 1956. Meaning in music. *Chicago: University of Chi-cago Press. Emotion and Meaning in Music*.

[Morris 1998] Morris, R. D. 1998. Voice-leading spaces. *Music Theory Spectrum* 20(2):175–208.

[Nakamura et al. 2015] Nakamura, E.; Cuvillier, P.; Cont, A.; Ono, N.; and Sagayama, S. 2015. Autoregressive hidden semi-markov model of symbolic music performance for score following. In *16th International Society for Music Information Retrieval Conference (ISMIR)*.

[Nettles and Graf 1997] Nettles, B., and Graf, R. 1997. *The chord scale theory & jazz harmony*. Advance music.

[Sheh and Ellis 2003] Sheh, A., and Ellis, D. P. 2003. Chord segmentation and recognition using em-trained hidden markov models. In *14th International Society for Music Information Retrieval Conference (ISMIR)*. Johns Hopkins University.

[Tymoczko 2004] Tymoczko, D. 2004. Scale networks and debussy. *Journal of Music Theory* 48(2):219–294.

[Wang and Dubnov 2014] Wang, C.-i., and Dubnov, S. 2014. Guided music synthesis with variable markov oracle. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[Wang and Dubnov 2015] Wang, C.-i., and Dubnov, S. 2015. The variable markov oracle: Algorithms for human gesture applications. *IEEE MultiMedia* 22(4):52–67.

[Wang, Hsu, and Dubnov 2016] Wang, C.-i.; Hsu, J.; and Dubnov, S. 2016. Machine improvisation with variable markov oracle: Toward guided and structured improvisation. *Computers in Entertainment (CIE)* 14(3):4.