Vol.2016-MUS-112 No.22
2016/8/1

# 音型の反復と変形に基づく階層ベイズ音楽言語モデルとMIDI演奏のリズム採譜への応用

中村 栄太[1,a)]　糸山 克寿[1]　吉井 和佳[1]

**概要**：本稿では，音型の反復構造を記述するベイジアン音楽言語モデルに基づくリズム採譜（即ち，MIDI演奏信号からの音価の自動認識）の手法について論ずる．自動採譜では従来，多数の楽譜データからの統計学習により構成した，音楽の一般的特徴を表す言語モデルが用いられてきた．一方で，多くの楽曲は反復構造を持ち，限られた音型から成っているため，楽曲ごとに個別の文法が学習できれば，より精密な言語モデルが得られると考えられる．ここで問題となるのは，演奏から間接的に得られる楽譜に対してその文法をどのように学習するかと，変形を含む音型の反復をどう扱うかである．本稿では，楽曲ごとに個別の言語モデルの生成を記述するディリクレ過程と変形を含む音型の組み合わせによる音符の生成過程を記述する階層 HMM（隠れマルコフモデル）を結合した階層ベイズモデルを提案する．このモデルに基づき，演奏信号から楽譜とその背後にある楽曲ごとに個別の言語モデルを同時に推定するための推論アルゴリズムを導出する．提案モデルにより従来の HMM よりもリズム採譜の精度が向上することを確認した．

## 1. はじめに

音楽音響信号から音高とリズムの情報を抽出する（自動）採譜は，音楽情報処理における根本的問題の一つである．これまで音高情報を抽出するために楽器音の音響モデルに関する研究が多く行われてきた [1,2]．一方で，音符の音価（即ち楽譜で記述される音の長さ）を認識する，リズム採譜に関する研究もされている [3–8]．これらの研究で，採譜には楽譜に関する事前知識を用いるのが重要であると分かってきており，この事前知識を表現する音楽言語モデルを構成するための機械学習手法が研究されてきた．中でも，楽譜データから音楽の特徴を統計学習するために，音声認識でも用いられる [9] HMM（隠れマルコフモデル）が広く用いられている [5–7, 10, 11]．これにより従来，多数の楽譜データにより学習することで，音楽の一般的特徴を表す言語モデル（一般的音楽モデルと呼ぶ）が構成されてきた．

多くの楽曲は反復構造を持つため [12–15]，個々の楽曲は音楽的に可能な音型（あるいは音符パターン）の中から選ばれた，限られたもので構成されている．この様に楽曲ごとに個別の「コンパクトな」文法を学習できれば，一般的音楽モデルより精密な言語モデルが得られると考えられる．反復構造が予め与えられれば個々の楽曲の文法は音型に基づく音符列の生成モデル [6, 8] を用いて学習できる．

[1] 京都大学
606-8501 京都府京都市左京区吉田本町
[a)] enakamura@sap.ist.i.kyoto-u.ac.jp

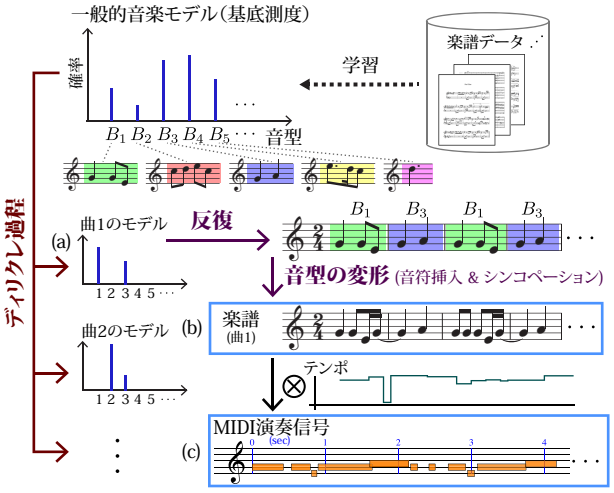


**図 1** 提案モデルの概要図．提案モデルは音型の不完全反復による楽譜生成を記述する言語モデルと時間変動を含む音楽演奏生成を記述する演奏モデルから成っている．

逆に，用いられている音型が分かれば，反復構造はより正確に推定できる．この鶏と卵の関係を解くためには，文法学習と構造推定を同時に行う枠組みが必要である．

この文法学習と構造推定の相互関係は，自然言語処理の分野でも話題となっており，ベイズ学習に基づく手法が開発されている [16, 17]．例えば，単語の切り出しと語彙の同時学習手法 [18] は，音符と文字，音型と単語の対応を考えると，今の問題と類似している．また，最近では単語の変形の問題も注目されている [19, 20]．反復パターンの変形のモデルは，音楽で一般的に現れる不完全反復 [21] を捉え

るために必要である．

本稿では，演奏信号から楽譜とその背後にある個別の言語モデルを同時学習するための階層ベイズモデルを提案する（図 1）．このモデルは，(a) 音型に基づく言語モデルの生成，(b) 言語モデルによる音符列の生成，(c) 音符列に基づく演奏信号の生成からなる，3 つの過程を階層的に記述する．(a) はいわば作曲者が楽曲で用いる音型を選択する過程をモデル化したもので，ディリクレ過程 [22] を用いて一般的音楽モデルから楽曲ごとの個別の言語モデルが生成される過程を記述する．(b) では，不完全反復を含む音符の生成を，音型の変形を記述する確率モデルにより表現する．演奏の生成過程は，楽譜・演奏マッチングの研究で従来用いられているテンポ変動のモデル [23, 24] により記述する．本稿では楽譜要素のうちリズムの側面のみ扱い，単旋律音楽に限定して議論することにする．

本研究の貢献は，ベイズモデルに基づくリズム採譜と背後にある文法・構造学習を同時に行う手法の開発と音型の不完全反復の取り扱いにある．音楽におけるベイジアン文法学習は，PCFG（確率的文脈自由文法）モデルに基づくものがあるが，変形を含まない音型を用いたもの [8] や音型を用いない音符単位の生成に関するもの [25, 26] に限られていた．自然言語処理分野でも，変形を含む文章に対する文法学習と単語の切り出しを同時に行う手法は提案されていない．提案モデルに基づくリズム採譜手法を評価し，従来の HMM より高精度でリズム採譜ができることを示す．

## 2. 関連研究

本節では，従来提案されてきた採譜のための統計モデルに基づく音楽言語モデルについてレビューする．これらのモデルは，提案モデルのベースになるもの，あるいは比較評価の対象となるものである．言語モデルの出力は音符列 $x_{1:N} = x_1 \cdots x_N$（$N$ は音符数）である．（以降，時系列に対して同様の記法を用いる．）本稿ではリズムの側面のみに注目しているので，$x_n$ は $n$ 番目の音符の音価を表す．音価は，楽譜に記される音長の全音符に対する相対的な値として定義する（例えば，四分音符は $x = 1/4$，付点二分音符は $x = 3/4$ など）．

### 2.1 音符マルコフモデル

音符列のマルコフモデルが初期の研究で提案されている（図 2）[6]．1 次のモデルでは，$x_{1:N}$ の確率は遷移確率の積として次のように与えられる．

$$P(x_{1:N}) = \prod_{n=1}^{N} P(x_n|x_{n-1}). \qquad (1)$$

ここで $P(x_1|x_0) \equiv P(x_1)$ は初期確率を表すものとする．同様に，$P(x_n|x_{n-1})$ を $P(x_n|x_{n-1}, \ldots, x_{n-p+1})$ に代えることで $p$ 次のマルコフモデルが定義できる．

この音符マルコフモデルには，音価列の論理制約に表現できないものがあるという欠点がある．例えば，三連符は 3 つ組あるいは 2 倍の三連符とのペアでのみ現れるという制約は，どれだけ高次のマルコフモデルを用いても表現できない．また拍節構造をモデルに組み込むこともできない．

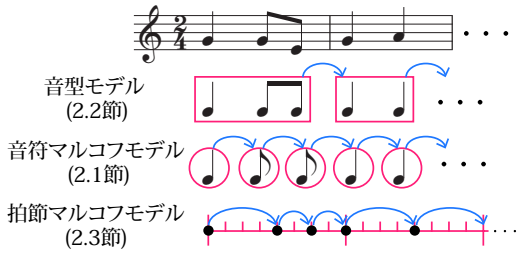


**図 2** 従来研究における 3 種類の楽譜の表現方法 [5–8].

### 2.2 音型モデル

クラシック音楽やポピュラー音楽などの多くの音楽では，楽譜は拍節構造を持つため，これを表現するモデルとして音型モデルが提案されている [6, 8]．音型マルコフモデル [6] では，小節などの単位ごとの音型の集合を状態空間として扱う．以下，音型の種類の数を $K$ とし，各音型を $B_k = z_{k,1} \cdots z_{k,L}$ $(k = 1, \ldots, K)$ と記す．ここで，$z_{k,\ell}$ $(\ell = 1, \ldots, L)$ は音型 $k$ の $\ell$ 番目の音符を表す．音型の系列 $w_1 \cdots w_I$ $(w_i \in \{B_k\}_{k=1}^K)$ の確率は，遷移確率 $\pi_{kk'} = P(w_i = B_{k'}|w_{i-1} = B_k)$ と初期確率 $\pi_k^0 = P(w_1 = B_k)$ の積で与えられる．

音符列（$z_{1:M}$ と記す）は生成された音型列 $w_{i:I}$ を結合することにより得られ，その確率は 2 層の階層マルコフモデル [27] により記述される．上の階層は音型，下の階層は音符に対応している．各音符 $z_m$ は $(k, \ell)$ のペアにより表され，遷移確率は次で与えられる．

$$P\big(z_m{=}(k', \ell')|z_{m-1}{=}(k, \ell)\big) = \delta_{\ell L}\pi_{kk'}\delta_{\ell' 1} + \delta_{kk'}\delta_{\ell'(\ell+1)}.$$

ここで，$\delta$ は Kronecker のデルタを表す．PCFG に基づく同様のモデルも提案されている [8]．

音型モデルのもう一つの長所は，三連符の論理制約などを記述できることである．一方で，シンコペーションの取扱いはこのモデルの問題点である．シンコペーションは，小節線など音型の境界をまたぐ音符を含むため，音型モデルでは簡単には記述できない．

### 2.3 拍節マルコフモデル

拍節構造を表現できるもう一つのモデルは，音符を小節内のビート位置で記述する，拍節（グリッド）マルコフモデルである [5, 7]．$n$ 番目の音符のビート位置を $s_n$ で表し，小節の長さを $G$ と記す．音符列の生成は，ビート位置のグリッド上のマルコフ過程により記述され，系列 $s_{1:N+1}$ の確率は遷移確率 $P(s_n|s_{n-1})$ の積により式 (1) と同様に与えられる．$n$ 番目の音符の音価は，

$$x_n = \begin{cases} s_{n+1} - s_n, & (s_{n+1} > s_n); \\ G + s_{n+1} - s_n, & (s_{n+1} \leq s_n), \end{cases} \qquad (2)$$

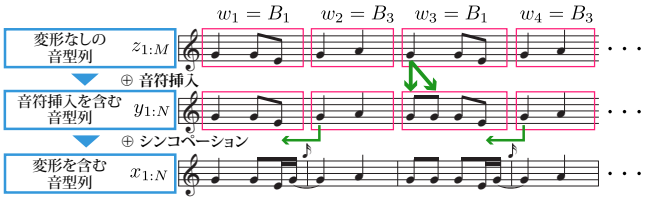


**図 3** 音型の変形モデル.

で与えられ，$s_{n+1}$ が $s_n$ よりも小さい場合は，次の小節のビート位置を表すものとする．

元々の拍節モデルでは，小節長までの音価しか記述していないが，より大きな音価を記述するようにモデルを拡張できる．このために，まず各音符 $n$ に対して，離散変数 $j_n = 0, 1, \ldots, J-1$ を導入し（$J$ は正の整数），$n-1$ 番目と $n$ 番目のオンセット位置の間に何小節飛び越えるかを表現する．つまり，$n$ 番目の音符の音価は $x_n = j_{n+1}G + s_{n+1} - s_n$ で与えられる．変数のペア $(s_n, j_n)$ を状態変数とすることにより，$JG$ までの音価を記述するモデルを構成できる．

拍節モデルの長所は，シンコペーションが扱えることである．式 (2) において $s_{n+1} \neq 0$ かつ $s_{n+1} < s_n$ の場合，$n$ 番目の音符はシンコペーションされた音符を表す．一方で，拍節モデルは新たな音価を加えた場合や異なる拍子に対して新たにモデル構成が必要など，モデルの拡張性の低さは短所といえる．

## 3. 提案モデル

提案モデルは，言語モデルと演奏モデルの 2 つの要素からなる．以下，これらのモデルの詳細と推論アルゴリズムについて述べる．

### 3.1 言語モデル

音型の不完全反復を含む楽曲構造を反映する音楽言語モデルを構成するため，2.2 節の音型モデルを次の 2 点で拡張する．まず音型モデルをディリクレ過程を用いてベイジアン拡張することで，限られた音型のみが選択されるコンパクトな文法を持ち，反復構造が誘導される言語モデルの生成過程をモデル化する．また音型の変形モデルを組込むことにより，不完全な反復を含む楽譜の生成過程をモデル化する．以下，変形モデルとベイジアン拡張の順に説明する．

#### 3.1.1 音型の変形

提案する言語モデルの出力 $x_{1:N}$ は，2.2 節の音型モデルにより生成された音符列 $z_{1:M}$ に変形を付すことで得られる（図 3）．本稿では実際の楽曲に現れる典型的な変形である，音符の挿入とシンコペーションに注目し，特にこれらの内で音価の総和が変化しないものを考える．

音符 $z_m$ から挿入により得られる音符列を $\tilde{z}_1, \ldots, \tilde{z}_Q$ と記すと，音符の挿入の確率は $P(\tilde{z}_1 \cdots \tilde{z}_Q | z_m)$ の形で表される．ここで，$\tilde{z}_1 + \cdots + \tilde{z}_Q = z_m$ が成立つ．記号 $Q$ により挿入後の音符数を表し，$q(= 1, \ldots, Q)$ を挿入後の音符の添字とする．また $C_h = \tilde{z}_1 \cdots \tilde{z}_Q$ $(h = 1, \ldots, H)$ により可能

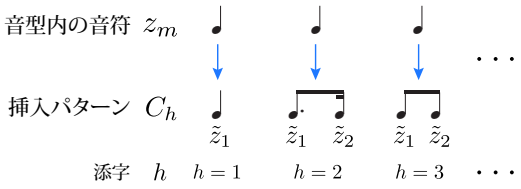


**図 4** 音符挿入による音型の変形の例.

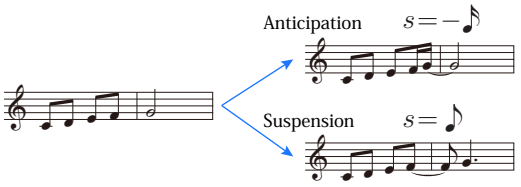


**図 5** シンコペーションによる音型の変形の例.

な挿入後の音符列の一つ（以下，挿入パターンと呼ぶ）を表し（図 4），確率 $P\big(C_h | z_m = (k, \ell)\big)$ を $\phi_{(k\ell)h}$ と記すことにする．音符列 $z_{1:M}$ 内の各音符 $z_m$ に対して挿入パターン $h_m$ を適用して得られる，音符列を $y_{1:N}$ で表す．この音符挿入による変形の過程は，2.2 節の音型モデルにさらに一つ下の階層のマルコフモデルを付け加えることで記述できる．

シンコペーションは音型の境界をまたいで，音型の最後の音符と次の音型の最初の音符の音価が同時に変形されたものと捉えられる．音符列 $y_{1:N}$ からシンコペーションを付すことで得られる音符列を $x_{1:N}$ と記すことにする．シンコペーションは，その度合を表す変数 $s$ によりパラメータ化でき，次の音符の変化に対応する（図 5）．

$$y_n \to x_n = y_n + s, \quad y_{n+1} \to x_{n+1} = y_{n+1} - s. \tag{3}$$

ここで $x_{n+1}$ はある音型の最初の音符を表す[*1]．変数 $s$ は正負両方の値を取ることができ，正の場合は掛留音（suspension），負の場合は先取音（anticipation）を表す．シンコペーションは，音型モデルの状態変数 $w_i$ を $(w_i, s_i)$ のペアに拡張することにより記述できる．$\theta_s = P(s)$ と記すと，遷移確率は $P(w_i = B_{k'}, s_i | w_{i-1} = B_k, s_{i-1}) = \pi_{kk'}\theta_{s_i}$ と拡張される．ここで単純化のため，変数ごとに確率が独立であると仮定した．初期確率の拡張も同様である．

以上をまとめると，言語モデルにより生成される楽譜 $x_{1:N}$ は，確率変数 $w_{1:I}$, $s_{1:N}$ と $h_{1:M}$ により明示され，各音符 $x_n$ は添字の組 $(k, \ell, h, q, s)$ により明示される．このモデルは次の遷移確率をもつマルコフモデルで記述される．

$$\begin{aligned} &P\big(x_n = (k', \ell', h', q', s') | x_{n-1} = (k, \ell, h, q, s)\big) \\ &\quad = \delta_{qQ}\phi_{(k'\ell')h'}\delta_{q'1}\big[\delta_{\ell L}\pi_{kk'}\theta_{s'}\delta_{\ell'1} + \delta_{kk'}\delta_{ss'}\delta_{(\ell+1)\ell'}\big] \\ &\qquad + \delta_{hh'}\delta_{(q+1)q'}\delta_{kk'}\delta_{\ell\ell'}\delta_{ss'}. \end{aligned} \tag{4}$$

#### 3.1.2 ディリクレ過程

言語モデルのパラメーター $\boldsymbol{\pi}_k = (\pi_{kk'})_{k'}$, $\boldsymbol{\pi}^0 = (\pi^0_k)_k$, $\boldsymbol{\phi}_{(k\ell)} = (\phi_{(k\ell)h})_h$ と $\boldsymbol{\theta} = (\theta_s)_s$ は音符列の統計的性質を表す．楽曲によって用いられる音型や変形の種類は異なるた

[*1] ここで導入した変数 $s$ は，2.3 節の $s_n$ とは無関係である．

め，個々の楽曲に対して異なるパラメータの値を考えることにする．ベイズモデルの枠組みでは，これらのパラメータは事前モデルから生成されるものとして記述される．音型の分布がスパースになるような事前モデルを用いれば，限られた音型のみから成るコンパクトな文法を記述できる．

生成される分布のスパースネスをコントロールできる事前モデルとしてディリクレ過程がある．有限次元の分布の場合，離散分布 $\boldsymbol{\pi}$ に対するディリクレ過程は，基底分布 $\boldsymbol{\omega}$ と集中度 $\alpha$ を用いて次のように記述される．

$$\boldsymbol{\pi} \sim \mathrm{DP}(\alpha, \boldsymbol{\omega}) = \mathsf{Dir}(\alpha\boldsymbol{\omega}). \tag{5}$$

ここで $\mathsf{Dir}(\cdot)$ はディリクレ分布を表す．このように生成された分布は $\mathbb{E}[\boldsymbol{\pi}] = \boldsymbol{\omega}$ を満たし，$\alpha$ が小さい場合は $\boldsymbol{\pi}$ の多くの要素は 0 となる傾向を持つ．

言語モデルの $\boldsymbol{\pi}_k$ と $\boldsymbol{\pi}^0$ に対してこのようなディリクレ事前モデルをおく．

$$\boldsymbol{\pi}_k \sim \mathsf{Dir}(\alpha\boldsymbol{\omega}_k), \quad \boldsymbol{\pi}^0 \sim \mathsf{Dir}(\alpha\boldsymbol{\omega}^0). \tag{6}$$

ハイパーパラメータ $\boldsymbol{\omega}_k$ と $\boldsymbol{\omega}^0$ に楽曲データから学習した一般的音楽モデル用いた場合，このディリクレ過程は作曲者が音楽的に可能な音型から個別の楽曲に使う音型を選択する過程と解釈できる．またこれらのハイパーパラメータを一様分布とすることにより，教師なし学習をすることもできる．集中度 $\alpha$ が小さい場合，各楽曲では小数の音型のみが使われることになり，コンパクトな文法が誘導される．$\boldsymbol{\phi}_{(k\ell)}$ と $\boldsymbol{\theta}$ に対しても同様にディリクレ事前分布をおく．

$$\boldsymbol{\phi}_{(k\ell)} \sim \mathsf{Dir}(\boldsymbol{\xi}), \quad \boldsymbol{\theta} \sim \mathsf{Dir}(\boldsymbol{\lambda}). \tag{7}$$

### 3.2 演奏モデル

リズム採譜の問題設定では，楽譜は演奏を通して間接的に得られるものである．MIDI 演奏信号は発音時刻の系列 $t_{1:N+1}$，あるいは同等である，IOI（発音時刻間隔）の系列 $d_{1:N}$ により明示される．ただし，IOI は $d_n = t_{n+1} - t_n$ で定義される．演奏モデルは，楽譜が与えられた時の確率 $P(d_{1:N}|x_{1:N})$ を記述するものである．

ここでは，演奏・楽譜マッチングの研究で提案されている線形動的システムに基づく演奏モデルを用いることにする [23,24]．これは，音楽演奏の時間揺らぎに対する次の 2 つの要因，演奏者の動作ノイズによる発音時刻の揺らぎとテンポ変動を確率的に記述したモデルである．テンポは，発音時刻の揺らぎを除いた時に IOI と音価の比 $d_n/x_n$ を表す潜在変数 $v_n$ により表され，その変動はマルコフ過程により記述される．テンポ変動と動作ノイズの両方をガウシアンで表すと，演奏モデルは次の式で与えられる．

$$v_n|v_{n-1} \sim \mathsf{N}(v_{n-1}, \sigma_v^2), \quad d_n|v_n, x_n \sim \mathsf{N}(v_n x_n, \sigma_t^2).$$

ここで $\sigma_v$ と $\sigma_t$ は，それぞれテンポ変動と動作ノイズの標準偏差を表す．演奏モデルの完全データ確率は次式で与えられる．

$$P(d_{1:N}, v_{1:N}|x_{1:N}) = \prod_{n=1}^{N} P(d_n|v_n, x_n)P(v_n|v_{n-1}). \tag{8}$$

ここで，$P(v_1|v_0) \equiv P(v_1)$ はテンポの初期確率を表す．実際上は，テンポ変数には音楽で典型的に用いられる範囲内で離散化した値を用いることで，推論が可能となる．式 (4) と (8) により，提案モデルは $Z_n \equiv (x_n, v_n)$ を潜在変数とする階層 HMM [27] として記述されることが分かる．

### 3.3 推論アルゴリズム

演奏信号 $D = d_{1:N}$ とハイパーパラメータ $\Lambda = (\boldsymbol{\omega}_k, \boldsymbol{\omega}^0, \alpha, \boldsymbol{\xi}, \boldsymbol{\lambda})$ が与えられた時に潜在変数 $Z = Z_{1:N}$ と言語モデルのパラメータ $\Theta = (\boldsymbol{\pi}_k, \boldsymbol{\pi}^0, \boldsymbol{\phi}_z, \boldsymbol{\theta})$ を同時推定することが目標である．モデル学習のステップでは，パラメータ $\Theta$ を事後確率 $P(\Theta|D,\Lambda)$ の最大化により推定する．採譜のステップでは，潜在変数を事後確率 $P(Z|D,\Theta) \propto P(Z,D|\Theta)$ の最大化により推定するが，これは標準的なビタビアルゴリズムを用いて行える．$P(\Theta|D,\Lambda)$ を直接最大化することは難しいが，ギブスサンプリングを用いて漸近的に厳密な推定を行える．これは同時確率 $P(Z,\Theta|D,\Lambda)$ からサンプリングを行うことで $P(\Theta|D,\Lambda)$ からのサンプルを得る方法である．

ギブスサンプリングは，$P(\Theta|Z,D,\Lambda)$ からのパラメータのサンプリングと $P(Z|\Theta,D,\Lambda)$ からの潜在変数のサンプリングを交互に行うものである．前者のサンプリングでは，モデルパラメータは事後ディリクレ分布からサンプルできる．例えば，遷移確率パラメータ $\boldsymbol{\pi}_k$ は

$$\boldsymbol{\pi}_k|Z,\Lambda \sim \mathsf{Dir}\big(\alpha\boldsymbol{\omega}_k + \boldsymbol{f}_k(Z)\big), \tag{9}$$

からサンプルできる（$f_{kk'}(Z)$ は $x_{1:N}$ の中で現れる遷移 $B_k \to B_{k'}$ の数，$\boldsymbol{f}_k(Z) = \big(f_{kk'}(Z)\big)_{k'}$ を表す）．他のパラメータのサンプリングも同様にできる．

確率 $P(Z|\Theta,D,\Lambda)$ からのサンプリングは，前向きフィルタリング・後向きサンプリングの方法により行える．前向きアルゴリズムにより前向き変数 $\alpha_n(Z_n) = P(Z_n, d_{1:n}|\Theta)$ を計算した後に，潜在変数は

$$P(Z_n|Z_{n+1:N}, D, \Theta) \propto \pi_{Z_n Z_{n+1}} P(d_{n+1}|Z_n, Z_{n+1})\alpha_n(Z_n)$$

により逐次サンプルできる．ただし，$Z_N$ は $P(Z_N|d_{1:N},\Theta) \propto \alpha_N(Z_N)$ からサンプルする．潜在変数である $x_{1:N}$ と $v_{1:N}$ は $P(Z|D,\Lambda)$ の中で強い相関を持つため，これらは同時にサンプルする必要があることに注意が必要である．

言語モデルの状態数は $\mathcal{O}(10^4)$ 程度になり得る上，テンポ変数との積空間はさらに多くの状態を持つため，前向きアルゴリズムの計算コストが実際的でないことがある．これは状態数が $N_\mathrm{s}$ の場合に，前向きアルゴリズムの計算量が $\mathcal{O}(NN_\mathrm{s}^2)$ であることから理解できる．この問題の解決案として，粒子フィルターを用いて前向き変数を近似的に

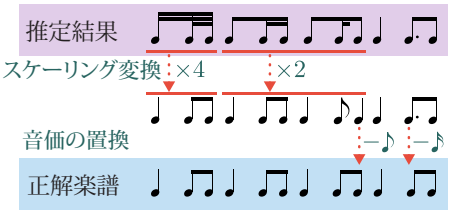


**図 6** リズム採譜結果の修正におけるスケーリング変換と音価置換の例.

計算することができる．これにより，粒子数を $N_\mathrm{p}$ とすると，計算量は $\mathcal{O}(NN_\mathrm{s}N_\mathrm{p})$ にまで削減できる．

## 4. 評価

### 4.1 セットアップ

提案モデルを評価するため，リズム採譜の精度を従来の HMM に基づくモデルと比較する．評価には，J-pop 30 曲のメロディーの MIDI 演奏を用いた（曲の長さは約 15 秒から 50 秒程度であった）．候補とする音価には，全音符から 16 分音符までの長さを持つ，通常の音符と付点音符と三連符を用いた．提案モデルにおける音型には，これらの音符からなる半音符分の長さの音型全てと全音符の長さを持つ (1)，(3/4,1/4) と (1/4,3/4) の三種類の音型を用いた．またこれらの音符のペアで表される，全ての音符の挿入パターンを用いた．シンコペーションの度合も，これらの音価とその負の値と 0 のどれかであるものとした．

提案モデルは，教師あり，半教師あり，教師なしの 3 つの学習条件と，音型の変形モデルありとなしの 2 通りの組合せで，全 6 通りの条件でテストした．教師あり学習で音型の変形なしの条件は，従来の音型 HMM [6] と同等である．この条件では，$(\boldsymbol{\pi}_k)_k$ と $\boldsymbol{\pi}^0$ を RWC データベース [28] のポピュラー曲 100 曲のメロディーを用いて学習した．半教師あり学習条件では，$(\boldsymbol{\omega}_k)_k$, $\boldsymbol{\omega}^0$ と $\boldsymbol{\lambda}$ を同じデータで学習した．教師なし学習条件では，$(\boldsymbol{\omega}_k)_k$ と $\boldsymbol{\omega}^0$ は一様分布とし，$\lambda_s$ は，$s=0$ ならば 10，それ以外の場合 0.05 とした．その他のハイパーパラメータは，$\alpha=700$, $\xi_h$ は挿入なしに対して 0.1，その他は 0.01，$\sigma_t=0.02$ 秒，そして $\sigma_v=0.06$ (秒/四分音符) とした [24]．

比較手法として，音符 HMM [6] と拍節 HMM [5]（それぞれ 1 次のもの）を実装した．これらのモデルは上述の RWC データベースを用いて学習した．また拍節 HMM では，小節単位を四分音符 2 個分（即ちポピュラー楽曲に多い 4/4 拍子では半小節分）とし，最大音価を音符 HMM や提案モデルと揃えるため $J=2$ とした．演奏モデルは，提案モデルと同じものを用いた．

リズム採譜の評価尺度として，「リズム修正率」を次の通り定義して用いた．まず，リズム修正コストを推定結果を正しい楽譜に修正するために必要な編集操作の最少数として定義し，これを音符数で割ったものをリズム修正率とした．リズムの修正に用いる編集操作として，音符単位の音価の置換のほか，音価列の部分列に対するスケール変換を用いた（図 6）．スケール変換を加えた理由は，音価の単位の選択には任意性があることによる．例えば，60 BPM のテンポで演奏した四分音符は，120 BPM で演奏した二分音符と同じ音長となる．詳細は紙面の都合上，他の論文に記すが，最小限必要な編集操作の数 $N_\mathrm{e}$ は Levenshtein 距離と同様な動的プログラミングによって計算できる．リズム修正率 $\mathscr{R}$ は $\mathscr{R}=N_\mathrm{e}/N$ で与えられ，小さいほど良い．

**表 1** リズム修正率 $\mathscr{R}$ の平均値と標準誤差．値が小さいほど良い．

| | モデル | 学習法 | 変形モデル | $\mathscr{R}$ [%] |
|---|---|---|---|---|
| 1 | 提案モデル | 教師なし | ✓ | $12.8 \pm 1.3$ |
| 2 | | 教師なし | | $16.5 \pm 1.8$ |
| 3 | | 半教師あり | ✓ | $\mathbf{6.6 \pm 1.0}$ |
| 4 | | 半教師あり | | $17.7 \pm 2.2$ |
| 5 | | 教師あり | ✓ | $7.8 \pm 1.2$ |
| 6 | | 教師あり | | $14.7 \pm 1.8$ |
| 7 | 音符 HMM [6] | 教師あり | | $7.9 \pm 1.4$ |
| 8 | 拍節 HMM [5] | 教師あり | | $7.3 \pm 1.3$ |

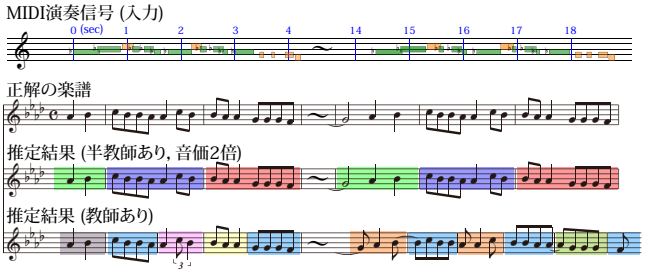


**図 7** 教師あり学習と半教師あり学習した提案モデル（ともに変形モデルあり）によるリズム採譜結果の比較.

### 4.2 結果

表 1 に結果を記す．提案モデルでは，音型の変形モデルを組込むことにより全ての学習条件で，リズム修正率の平均が向上した．変形ありの場合の教師あり学習と半教師あり学習の結果を比較すると，個々の楽曲の文法学習が有効的に働いていることが確認できる．また，教師なし条件では教師あり条件に比べて 5 ポイント精度が低かった．提案モデルで最も精度が高かった，変形ありの半教師あり学習の場合の結果は，従来の HMM に基づく手法よりも精度が高かった．

図 7 に示す結果の例では，教師あり学習した音型モデルに比べ，半教師あり学習した提案モデルにより反復構造が捉えられているとこが確認できる．教師あり学習の場合に推定誤りがあった箇所が，半教師あり学習の場合では反復構造と共に正しく推定されている．図 8 に示す結果の例では，提案モデルによってシンコペーションと音符の挿入による変形を含む音型の不完全反復構造が捉えられていることが確認できる．

提案法の性能を向上するための改良の余地は多い．一つ

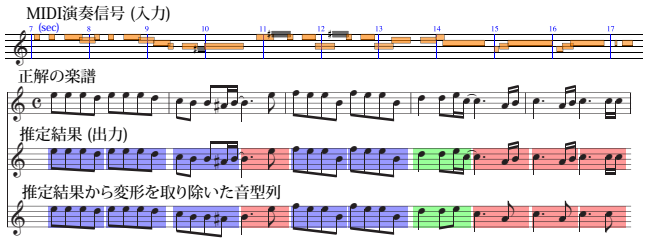


**図 8** 提案法（半教師あり，変形モデルあり）によるリズム採譜結果の例．上から 3 番目は推定結果，4 番目は変形を除いた推定結果を示している．

は可変長の音型モデルにより，音型の長さの選択に関する任意性を取り除くことである．これは階層ディリクレ過程を用いて行えると考えられる．また階層的な反復構造のモデル化および音高情報を用いた反復構造の推定も有効だと考えられる．

## 5. おわりに

音楽 MIDI 演奏信号から，楽譜とその背後にある楽曲ごとに個別の言語モデルを同時に推定する枠組みについて論じた．提案モデルにより，楽曲ごとのコンパクトな文法の学習および代表的な音型によるセグメンテーションができることが確認できた．反復構造に基づく文法と構造の同時学習の枠組みは，音響信号を入力とした現実的な採譜でも有効であると考えられる．今後，無限語彙モデルに拡張し，可変長の音型の自動学習を行う予定である．また多声音楽に対する拡張のため，声部構造の扱いも今後の課題である．

## 謝辞

有益な議論をして頂いた持橋大地氏と錦見亮氏に感謝する．本研究の一部は JST OngaCREST プロジェクト，JSPS 科研費 24220006, 26700020, 26280089, 15K16054, 16H01744, 16J05486 と栢森情報科学振興財団助成金による支援を受けて行われた．著者の中村は，日本学術振興会の特別研究員（PD）制度より支援を受けた．

### 参考文献

[1] A. Klapuri and M. Davy (eds.), *Signal Processing Methods for Music Transcription*, New York: Springer, 2006.
[2] E. Benetos *et al.*, “Automatic Music Transcription: Challenges and Future Directions,” *J. Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.
[3] A. T. Cemgil *et al.*, “Rhythm Quantization for Transcription,” *Comp. Music J.*, vol. 24, no. 2, pp. 60–76, 2000.
[4] D. Temperley, *The Cognition of Basic Musical Structures*, The MIT Press, 2001.
[5] C. Raphael, “Automated Rhythm Transcription,” *Proc. ISMIR*, pp. 99–107, 2001.
[6] H. Takeda *et al.*, “Hidden Markov Model for Automatic Transcription of MIDI Signals,” *Proc. MMSP*, pp. 428–431, 2002.
[7] M. Hamanaka *et al.*, “A Learning-Based Quantization: Unsupervised Estimation of the Model Parameters,” *Proc. ICMC*, pp. 369–372, 2003.
[8] M. Tsuchiya *et al.*, “Probabilistic Model of Two-Dimensional Rhythm Tree Structure Representation for Automatic Transcription of Polyphonic MIDI Signals,” *Proc. APSIPA*, paper id 14002890, pp. 1–6, 2013.
[9] S. Levinson *et al.*, “An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition,” *The Bell Sys. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.
[10] C. Ames, “The Markov Process as a Compositional Model: A Survey and Tutorial,’ *Leonardo*, vol. 22, no. 2, pp. 175–187, 1989.
[11] F. Pachet *et al.*, “Finite-Length Markov Processes with Constraints,” *Proc. IJCAI*, pp. 635–642, 2011.
[12] M. Cooper *et al.*, “Automatic Music Summarization via Similarity Analysis,” *Proc. ISMIR*, pp. 81–85, 2002.
[13] J. Paulus *et al.*, “State of the Art Report: Audio-Based Music Structure Analysis,” *Proc. ISMIR*, pp. 625–636, 2010.
[14] D. Meredith *et al.*, “Algorithms for Discovering Repeated Patterns in Multidimensional Representations of Polyphonic Music,” *J. New Music Res.*, vol. 31, no. 4, pp. 321–345, 2002.
[15] C. Wang *et al.*, “Music Pattern Discovery with Variable Markov Oracle: A Unified Approach to Symbolic and Audio Representations,” *Proc. ISMIR*, pp. 176–182, 2015.
[16] M. Johnson *et al.*, “Bayesian Inference for PCFGs via Markov Chain Monte Carlo,” *Proc. HLT-NAACL*, pp. 139–146, 2007.
[17] H. Shindo *et al.*, “Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing,” *Proc. ACL*, pp. 440–448, 2012.
[18] D. Mochihashi *et al.*, “Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling,” *Proc. ACL-IJCNPL*, pp. 100–108, 2009.
[19] T. Nakamura *et al.*, “Multimodal Concept and Word Learning Using Phoneme Sequences with Errors,” *Proc. IROS*, pp. 157–162, 2013.
[20] Y. Yang *et al.*, “A Log-Linear Model for Unsupervised Text Normalization,” *Proc. EMNLP*, pp. 61–72, 2013.
[21] L. Stein, *Structure & Style: The Study and Analysis of Musical Forms*, Summy-Birchard Inc., 1979.
[22] M. Jordan, “Dirichlet Processes, Chinese Restaurant Processes and All That,” Tutorial presentation at the NIPS Conference, 2005.
[23] C. Raphael, “Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models,” *IEEE TPAMI*, vol. 21, no. 4, pp. 360–370, 1999.
[24] E. Nakamura *et al.*, “A Stochastic Temporal Model of Polyphonic MIDI Performance with Ornaments,” *J. New Music Res.*, vol. 44, no. 4, pp. 287–304, 2015.
[25] M. Nakano *et al.*, “Bayesian Nonparametric Music Parser,” *Proc. ICASSP*, pp. 461–464, 2012.
[26] E. Nakamura *et al.*, “Tree-Structured Probabilistic Model of Monophonic Written Music Based on the Generative Theory of Tonal Music,” *Proc. ICASSP*, pp. 276–280, 2016.
[27] S. Fine *et al.*, “The Hierarchical Hidden Markov Model: Analysis and Applications,” *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.
[28] M. Goto *et al.*, “RWC Music Database: Popular, Classical, and Jazz Music Databases,” *Proc. ISMIR*, pp. 287–288, 2002.