Vol.2014-MUS-103 No.23
2014/5/24

潜在共通構造モデルに基づく音響信号間アライメント

前澤 陽[1,2,a)]　糸山 克寿[1]　吉井 和佳[1]　奥乃 博[1]

概要：本稿では，同一楽曲を演奏した複数の音響信号に対して時間軸対応付け（音響信号間アライメント）を行うための確率モデルを提案する．我々は，アライメント結果に基づいて演奏分析を行う応用を考えると，複数の演奏の背後に存在する潜在的な共通構造と各演奏に固有の時間的ゆらぎとを区別することが重要であると考えている．従来は，動的時間伸縮法（DTW）や Left-to-Right 型隠れマルコフモデル（LRHMM）を用いて，表層的な音響的類似度に基づいて対応点を探す手法が主流であった．一方，本研究では，複数の演奏に共通な状態系列を生成する上位 HMM と，上位 HMM で定められた順序で状態を遷移する演奏ごとに独立な下位 LRHMM を考え，両者を階層 HMM として確率的に統合する．このとき，上位 HMM においては，楽曲中で繰り返し登場する音響的特徴が同じ状態に割り当てられているので，楽曲自体の音楽構造の解析が容易に行える．さらに，下位 LRHMM においては，各状態での滞留時間に着目することで，各演奏に固有の時間的ゆらぎを調査することができる．実験の結果，音響信号間アライメント精度の点で，提案手法は従来法より優れていることが分かった．

1. はじめに

同じ曲を別々の人間が弾いた音響信号や，同じ詩を別々の人間が朗読した音声には，演奏者・朗読者の解釈の違いが反映される．このような解釈の違いを，計算機を用いて，可視化 [1,2]，検索 [3]，またはシームレスに再生 [4] することで，ユーザが，解釈の違いに対する理解を深めたり，好みの演奏者を探すことが可能になると期待される．このような応用では，複数の音響信号間における時間軸対応付け（音響信号間アライメント）が重要になる．

アライメントが幅広い環境・問題定義で動作するには，音響信号間の変動要因を，確率的生成モデルでモデル化することが重要である．なぜならば，確率的生成モデルは，音響信号に含まれる不確定要素を，明示的に表現できるためである．例えば，楽譜対音響信号アライメントの文脈では，演奏ミスが発生する過程をモデル化することで，演奏ミスに対して頑健なアライメント手法が確立している [5]．また，音響信号間アライメントでは，複数のパートが混合される過程をモデル化することで，楽曲を演奏するパート数の違いに対して頑健なアライメントを実現している [6]．

生成モデルとしての音響信号アライメントでは，音楽というメディアを奏でた音響信号の生成過程をモデル化する．そのため，音楽の生成過程と，音楽を演奏する音響信号の生成過程をモデル化する必要がある．そこで，アライメント手法には，次の二つの要件があると考えられる：

(1) 音楽の生成過程，とりわけ，楽曲に内在する構造をモデル化すること．西洋音楽を始めとする多くの音楽は，少数の，「動機」と呼ばれる短い音列を，繰り返したり変形させることで，大きな楽曲を構築する．つまり，音楽は，コンパクトな和音列の変形や反復として，上手く説明できると考えられる．本稿では，このように，楽曲を構成するコンパクトな構造を「**潜在共通構造**」と呼ぶ．一般的に，よりシンプルな構造を持つ生成モデルであるほど，頑健に動作する．そのため，頑健性を確保する上で，良い潜在共通構造のモデル化が重要になると考えらえる．

(2) 同一楽曲を演奏する過程をモデル化すること．具体的には，楽曲を説明する生成モデルパラメータの時系列は，全ての音響信号に対して，同じ順序で遷移すること．なぜならば，同一楽曲を演奏する音響信号は，背後に共通のシンボル列（楽譜等）を持っているためである．これを「**状態遷移順序の同一性**」と呼ぶ．状態遷移順序が同一であるならば，モデルパラメータが変化した地点を複数の音響信号で対応付けることにより，アライメントが求まる．

従来のアライメント手法では，楽曲を，ある決まった順序で遷移するよう制約された状態系列（Left-to-right Hidden Markov Model（HMM）；LRHMM）としてモデル化していた [6,7]．このようなモデルは，状態遷移順序の同一性は満たすものの，楽曲に内在する反復構造や，音符列の繰り返しといった要素は考慮しないため，潜在共通構造のモデルとして不適切である．そのため，LRHMM では，楽曲に内在する共通のパターンを認識できず，これらの手法が良好に動作するためには，瞬時的な音の生成モデルを別の手法で設定する必要があった [7]．

そこで，本稿では，潜在共通構造をモデル化しつつ，状態遷移順序の同一性を満たすような，音響信号アライメント手法を提案する．具体的には，潜在共通構造を表現するため，Ergodic HMM から適当な状態系列を生成し，状態

[1] 京都大学大学院情報学研究科
Kyoto University, Yoshida Honmachi, Sakyo, Kyoto, 606–8501, Japan
[2] ヤマハ株式会社
Yamaha Corporation
[a)] akira.maezawa@gmail.com

遷移順序の同一性を満たすため，Ergodic HMM から生成された状態系列の順序で，個々の楽曲の状態が遷移するような階層モデルを考える．なお，本稿では主に音楽音響信号を対象として扱うが，音声信号も提案手法で同様に扱えることに留意されたい．

2. 関連研究

アライメントには，楽譜表現といったシンボル列を事前に与える必要があるもの [8,9] と，そうでないものに分けられる．本稿では，より幅広い楽曲や音声に対処するため，後者を考える．

前述した生成モデルアプローチの他に，アライメントは，「音楽的に似た二つの音」の距離が小さくなるような距離尺度を設計した上で，音響信号間の各時刻の対における距離行列上の経路探索問題や [10, 11]，連続的な経路の事後分布推定問題 [12] として定式化することができる．このような方法は，「似た音」を距離尺度として適切に設計できれば良好に動作するが，単一の尺度として適切に表現できない場合，このような手法は適用できない．一方で，単一の距離尺度として記述するには，限界があるような複雑なものも，生成モデルとして記述することでモデル化できる．例えば，演奏ミス [5]，楽器間における音色や音量バランスの違い [13–15]，演奏されるパートの違い [6] といった違いは，生成モデルではシンプルに記述できるが，距離尺度として設計するのは困難であると考えられる．

なお，本稿で考える，楽曲の反復箇所をグルーピングするようなコンパクトな表現（潜在共通構造）は，楽曲構造解析では，広く用いられているアイディア [16] である．

3. 定式化

本手法は，同一楽曲を演奏した D 個の音響信号から抽出した，D 組の特徴量系列を与えられた時，潜在共通構造を明示的に扱いながら，状態遷移順序の同一性を満たすアライメント手法である．なお，d 番目の音響信号は，特徴系列の長さが T_d であり，特徴量時系列を $x(d, 1 \cdots T_d)$ とする．

本手法のポイントは，瞬時的な音響特徴量の生成過程を，コンパクトな状態空間モデル（Ergodic HMM）で学習することに加え，このような生成過程から仮想的に生成された系列に対して，音響信号同士のアライメントを算出することである．ここでの「仮想的に生成された系列」とは，楽曲を，コンパクトな状態空間モデルで説明した状態系列である．コンパクトな状態空間モデルに対して音響特徴量の生成モデルを学習するため，楽曲を通して繰り返し出現する似た音響特徴量は，一つの生成過程として学習される．そのため，一つの生成過程を学習する時に，より多くのデータを用いることが可能となり，頑健な学習が実現できると期待される．

図 1 に概念図を示す．この図では，楽曲は 3 つの状態 A, B, C から構成され，「ABCB」という順序で楽曲が進行する．「楽曲進行」とは，任意の長さに固定した Ergodic HMM の系列である．この場合，楽曲進行は長さ 5 の系列 $S_1 \cdots S_5$ であり，S_1 が状態 A, S_2 と S_3 が状態 B，S_4 が状態 C，S_5 が状態 B に割り当てられている．楽曲進行を，S_1 から S_5 まで順に辿ると，楽曲を表す系列「ABCB」が再現される．また，音響信号 1 から D では，楽曲進行 $S_1 \cdots S_5$ に対するアライメントを求めている．これにより，楽曲自身の進行「ABCB」に対するアライメントを，間接的に求めていることが分かる．単一の系列である楽曲進行に対してアライメントを行うため，状態遷移順序の同一性が満たされることが分かる．また，状態 A，B，C の生成過程は，各音響信号のアライメントと，楽曲進行を基に，状態 A，B，C にあるフレームをそれぞれ抽出し，抽出されたフレームに基づいて学習を行うことができる．例えば，状態 B を学習するには，共通状態系列が S_2，S_3 及び S_5 であるフレームを音響信号 1 から D の間で抽出し，抽出されたデータから学習を行えばよい．従来 [7] では，共通状態一つ一つに対して，個別の生成過程を割り当てていたため，この

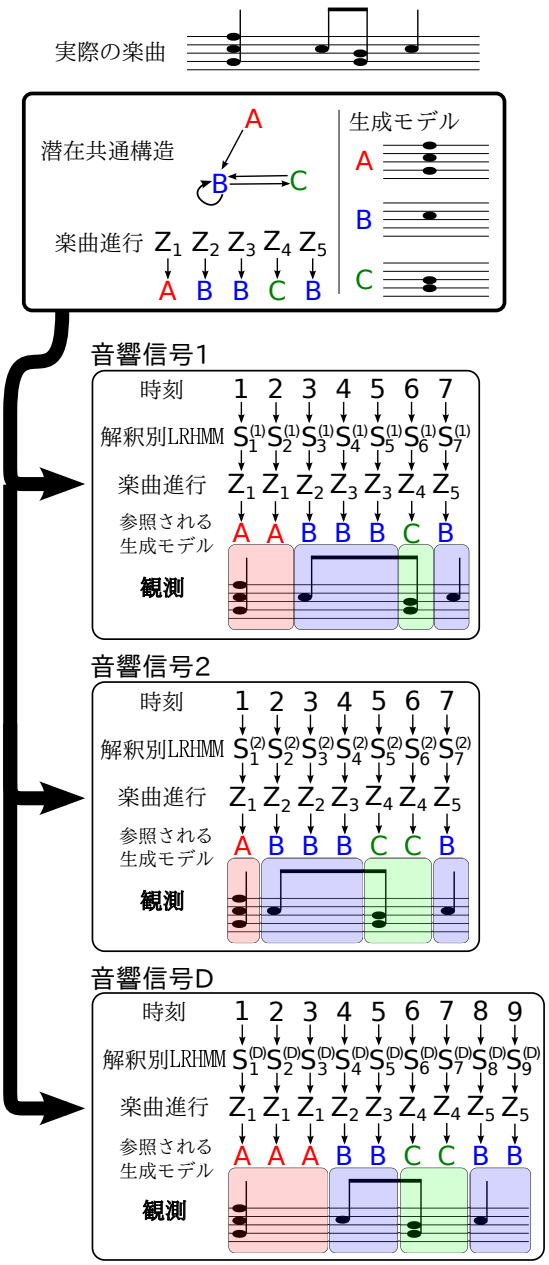

図 1 本手法のコンセプト．本手法では，共通構造を持ち，状態遷移順序の同一性を持つ複数の音響信号を，共通構造から生成された，状態系列の順序に従う状態系列という階層構造としてモデル化する．

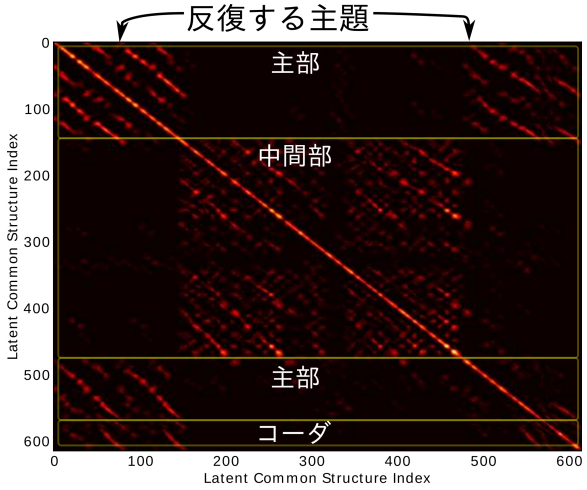

図 2 Chopin Op. 41-2 の潜在共通構造から類似行列を求めたもの．分かりやすさのため，類似行列上に，楽曲構造に対応する枠を表示している．

ケースでは，S_1 から S_5 の 5 状態の生成過程を学習していた．つまり，S_2，S_3 と S_5 が同一であるいった，楽曲に内在する冗長性を反映することができなかった．

次に，このような概念を，確率的生成モデルとして記述することを考える．具体的には、潜在共通構造を，状態数 S の HMM として記述し，その HMM から，長さ N の状態系列を生成する．このような HMM を「潜在共通構造 HMM」，潜在構造 HMM から生成された状態系列を「楽曲進行」と呼ぶ．また，それぞれの音響信号は，状態数 N の LRHMM としてモデル化し，n 番目の**状態**を，楽曲構造における，n 番目の状態に割り当てられている観測モデルに結びつける．このような LRHMM を「解釈別 LRHMM」と呼ぶ．このように，潜在的共通構造から，単一の状態系列を一旦生成し，その状態系列の順序で遷移するような LRHMM を置くことで，複数の音響信号間の状態遷移の同一性を保証する．

3.1 潜在共通構造 HMM と楽曲進行のモデル化

楽曲進行は，前述の通り，状態数 S の潜在共通構造 HMM から生成された，長さ N の状態系列 $Z(n = 1 \cdots N)$ として表す．これは，一つの楽曲は，最大 S 種類の音から構成され，最大 N 回音の変化が生じると仮定することに相当する．ここで，楽曲進行 $Z(n)$ を one-of-S の二値変数として表現する．つまり，$Z(n)$ を S 次元の二値ベクトルとして表し，$Z(n)$ の状態が s であるとき，$Z_s(n) = 1$ であり，それ以外の要素を 0 にする。これを踏まえ，$Z(1 \cdots N)$ を，初期状態 π，状態遷移確率 τ に従う Ergodic HMM（任意の状態間を遷移できる HMM）として定義する:

$$p(Z|\pi,\tau) = \prod_{s=1}^{S} \pi_s^{Z_s(1)} \prod_{n=2,s'=1,s=1}^{N,S,S} \tau_s(s')^{Z_{s'}(n-1)Z_s(n)} \quad (1)$$

潜在共通構造 HMM の各状態 s には，短時間の音響信号が生成される過程におけるパラメータ $\theta(s)$ が割り当てられている．$\theta(s)$ とは，たとえば，楽曲であれば音高の組み合わせ（「ド＋ミ＋ファ」），発話であれば音素に相当するパラメータなどが割り当てられる．なお，状態遷移 $\tau(s)$ に対しては $p(\tau(s)|\tau_0) = \mathrm{Dir}(\tau|\tau_0)$ といった，ディリクレ分布を事前分布として設定し，また，初期状態確率 π に対しては $p(\pi|\pi_0) = \mathrm{Dir}(\pi|\pi_0)$ を割り当てる．

本手法の潜在共通構造にはどのような情報が含まれるのだろうか．**図 2** に，Chopin Op. 41-2 の潜在共通構造から類似行列を求めたものを図示する [*1]．この曲は三部形式であり，ニ長調の主部では，ABAC という形の主題が 2 回繰り返され，ロ長調の中間部を経た後，主部の主題が 1 回再生されたあと，主題に出現するモチーフ “C” に基づく短いコーダで終わる．一方，類似行列を分析すると，大まかなチェッカーボード状のパターンから，三部形式であることが分かる．また，主部に対応する区間の対角線から，主題が ABAC の形であり，左下の対角線に対して 3 本のほぼ平行な線が生じていることなどから，モチーフの繰り返しが見て取れる． このことから，本手法の潜在共通構造は，楽曲の構造を示すような情報が含まれることが示唆される．

このような構造が見えてくるのは，潜在共通構造を ergodic HMM としてモデル化しているからである．従来用いられている LRHMM [7] では，同じ状態に戻らないため，状態系列の類似性から楽曲の特徴を掴むことはできない．

3.2 解釈別 LRHMM のモデル化

次に，各音響信号に対して，状態遷移の同一性を満たしながらも，それぞれ各状態に停留する時間が独立になるようなモデルを考える．そこで，それぞれの音響信号に割り当てらた解釈別 LRHMM を，楽曲進行の進み具合を示すようにする．LRHMM は，状態 n が，状態 n か $n+1$ にしか遷移できず，かつ，時系列の始点と終点を任意の状態に拘束したものであるため，状態遷移順序の同一性を保証する．そこで，d 番目の音響信号における状態系列 $S^{(d)}$ を，隣接する状態への遷移確率を $\eta(d,n)$ とした，状態数 N の LRHMM としてモデル化する：

$$\begin{aligned} p(S^{(d)}(t = 1 \cdots T_d)) &= \delta(s,1)^{S_s(d,1)} \delta(s,S)^{S_s(d,T_d)} \\ &\times \prod_{t=1,n}^{T_d,N} \Big[\eta(d,n)^{S_n(d,t-1)S_{n+1}(d,t)} \\ &\times (1-\eta(d,n))^{S_n(d,t-1)S_n(d,t)} \Big] \end{aligned} \quad (2)$$

ここで、$\delta(x,y)$ とは，Kronecker Delta であり，$x = y$ の時のみ 1 であり，それ以外は 0 であるような関数である．また，隣接する状態に遷移する確率 $\eta(d,n)$ に対して，事前分布 $p(\eta(d,n)|a_0,b_0) = \mathrm{Beta}(\eta(d,n)|a_0,b_0)$ を割り当てる．

図 3 に，潜在共通構造 HMM の状態が変化したと判定された地点の例を図示する．この図から，アタックのように，楽器の音において特徴量が大きく変化する区間で，状態が切り替わることが分かる．解釈別 LRHMM は，状態遷移順序の同一性を満たしているため，図の縦線の数は，全ての音響信号で同じである．そのため，i 番目に出現する縦線の位置を，全ての i に対して対応付けることで，アライメントが求まる．

[*1] 類似行列 $R(i,j)$ を求めるには，楽曲進行から自己遷移を省いた系列 $Z'(n)$ を求め，$R(i,j) = \delta(Z'(i),Z'(j))$ とした．次に，$R(i,j)$ を 2 次元画像としてみなし，対角線を強調するようなフィルタを畳み込んだ．

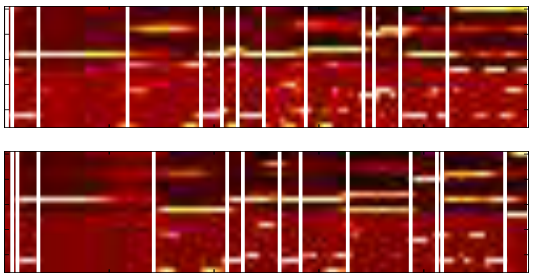

図 3 二つの演奏に対する特徴量の時系列に対して，潜在共通構造の状態変化地点に縦線を引いた図（縦軸＝特徴量次元，横軸＝時間）．特徴量の傾向が変わる地点と，潜在共通構造の状態変化地点が対応づいていることが分かる．なお，特徴量は Chroma ベクトルと Δchroma を用いてる．

3.3 音響信号の生成過程

上記の議論を踏まえると，音響信号 d の時刻 t での観測尤度は，まず，解釈別 LRHMM から，時刻 t の楽曲進行を割り出し，次に，時刻 t の楽曲進行に対応する状態に対応する観測尤度を，潜在共通構造 HMM から参照すればよい：

$$p(x(d,t)|Z,S^{(d)}) = \prod_{s,n} p(x(d,t)|\theta(s))^{Z_s(n)S_n^{(d)}(t)} \quad (3)$$

ここで，$p(x|\theta(s))$ とは，状態 s において特徴量 x を観測する尤度であり，そのパラメータ $\theta(s)$ とは，事前分布 $p(\theta(s)|\theta_0)$ から生成されたものとする．

ここで重要なのは，上のモデルについて，データを表現するのに十分な生成過程を置くことができることである．本稿では簡単のため，$p(x(d,t)|\theta(s))$ を $U=\dim(x)$ 次元の Normal 分布とし，$p(\theta(s)|\theta_0)$ を U 次元の Normal-Gamma 分布とする．つまり，$\theta_s \in \{\mu_s, \lambda_s\}$, $\theta_0 \in \{m_0, \nu_0, a_0, b_0\}$ とし，$p(\mu_s,\lambda_s|m_0,\nu_0,a_0,b_0) \propto \lambda_s^{a_0-\frac{1}{2}} e^{-(\mu_s-m_0)^2\lambda_s\nu_0 - b_0\lambda_s}$ とする．しかし，節 2 で紹介したような生成モデルを活用することで，距離尺度として表現が難しい，幅広い音のバリエーションをモデル化できることに留意されたい．

3.4 事後分布の推定

このモデルは共役な指数分布族から構築される．そのため，変分ベイズ法を用いることで，事後分布を解析的に推定することが可能である．紙面の制約上，詳しい導出は省略するが，まず，潜在共通構造 HMM，解釈別 LRHMM，各生成過程の独立性を仮定する．すると，潜在共通構造 HMM は，HMM の前向き後ろ向きアルゴリズムを用いて期待値を計算ができる．この時，楽曲進行の位置 n に対する，状態 s の出力確率 $O_s^{(Z)}(n)$ と状態遷移確率 $T_{s,s'}^{(Z)}(n)$ は，次のように与えられる：

$$O_s^{(Z)}(n) = \exp\Big(\sum_{d,t}\langle S_n(d,t)\rangle\langle\log p(x(d,t)|\theta(s))\rangle\Big) \quad (4)$$

$$T_{s,s'}^{(Z)} = \exp\langle\log\tau_{s'}(s)\rangle \quad (5)$$

解釈別 LRHMM も，前向き後ろ向きアルゴリズムで，期待値を求めることができる．このとき，音響信号 d の時刻 t における，楽曲進行 n の出力確率 $O_n^{(S)}(d,t)$ と，遷移確率 $T_{n,n'}^{(S)}(d)$ は次のように与えられる：

$$O_n^{(S)}(d,t) = \exp\Big(\sum_s\langle Z_s(n)\rangle\langle\log p(x(d,t)|\theta(s))\rangle\Big) \quad (6)$$

$$T_{n,n'}^{(S)}(d) = \begin{cases}\exp\langle\log\eta(d,n)\rangle & n=n'\\ \exp\langle\log(1-\eta(d,n))\rangle & n+1=n'\end{cases}$$

また，状態 s の生成過程は，データの対数観測尤度が $\sum_t\sum_n\langle Z_n(s)\rangle\langle S_n(d)(t)\rangle\log p(x(d,t)|\theta(s))$ であった場合の事後分布を求めればよい．例えば，今回のように $p(x(d,t)|\theta_s)$ を正規分布とし，$p(\theta_s|\theta_0)$ を Normal-Gamma 分布とした場合，事後分布は次のように更新される：

$$q(\mu_s,\lambda_s) = \mathcal{NG}(m_s,\nu_s,a_s,b_s) \quad (7)$$

ただし，m_s，ν_s，a_s，b_s はそれぞれ次のように与えられる：

$$\bar{N}_s = \sum_{d,n,t}\langle Z_s(n)\rangle\langle S_n^{(d)}(t)\rangle$$

$$\bar{\mu}_s = \frac{1}{\bar{N}_s}\sum_{d,n,t}\langle Z_s(n)\rangle\langle S_n^{(d)}(t)\rangle x(d,t)$$

$$\bar{\Sigma}_s = \frac{1}{\bar{N}_s}\sum_{d,n,t}\langle Z_s(n)\rangle\langle S_n^{(d)}(t)\rangle(x(d,t)-\mu_s)^2$$

$$\nu_s = \nu_0+\bar{N}_s;\ m_s = \frac{\nu_0 m_0+\bar{N}_s\bar{\mu}_s}{\nu_0+\bar{N}_s}$$

$$a_s = a_0+\frac{1}{2}\bar{N}_s;\ b_s = b_0+\frac{1}{2}\Big(\bar{N}_s\bar{\Sigma}_s+\frac{\nu_0\bar{N}_s}{\nu_0+\bar{N}_s}(\bar{\mu}_s-m_0)^2\Big)$$

4. 潜在共通構造モデルの改良

潜在共通構造に基づくアライメント手法は，ここまで紹介したように，HMM を階層的にモデル化することで，潜在共通構造をモデル化しながら，状態遷移順序の同一性を満たす．しかし，時系列モデルとして楽曲構造に HMM より柔軟なモデルを用いたり，解釈別 LRHMM に LRHMM より柔軟なモデルを用いることで，より精度が向上する可能性がある．そこで，以下では，潜在共通構造 HMM および，解釈別 LRHMM の改良を二つ提案する．

4.1 潜在共通構造 HMM の状態数を無限化する

現在の解釈共通構造 HMM では，楽曲の表現に用いる状態数 S を任意の整数にしているため，設定する S の値によって推定結果が変わる可能性がある．このような問題に対処するため，ディリクレ過程 [17] を用いて，S を無限にさせながら適当な縮退効果を持たせることで，用いられる実効的な状態数が，データの複雑に応じて増減するといったことも可能である．そこで，HMM の状態をディリクレ過程としてモデル化した，Nonparametric Bayesian HMM [18] を用いて，S の実質的な数をデータの複雑さに応じて変えることを考える．

Nonparametric HMM を適用する際，潜在共通構造の状態遷移確率 $\tau(m)$ の事前分布は，次のように変わる：

$$\tau(m) \sim \mathrm{GEM}(\alpha) \quad (8)$$

ここで，GEM(α) とは，次のような確率過程に従って生成される変数のことを指す：

$$w_i \sim \mathrm{Beta}(1,\alpha);\ \tau_{m'}(m) = \prod_{i=1}^{m'-1}(1-w_i)w_{m'}$$

つまり，τ_i は，長さ 1 の棒を与え，それをおおよそ 1 対 α の比率で折り，その右半分を取る，というプロセスを i 回繰り返した時，最後のステップで得られた，左半分の棒の長さを割り当てることに相当する．

4.2 解釈特有 LRHMM のセミマルコフ化

楽曲進行において任意の状態に留まる時間は，演奏による違いがあるものの，楽曲に記載されているテンポ指示によって大まかに定められる．しかし，節 3 で定式化した，解釈別 LRHMM では，ある状態に留まる時間が幾何分布に従うことを暗に仮定している．そのため，状態停留時間が，特定の時間に集中する傾向を，表現できない．

そこで，LRHMM の状態遷移に明示的な継続長を持たせることを考える．ここで，各状態 n の停留時間を，全ての楽曲で共通の正規分布で表すとする：

$$p(l|\mu(n),c) = \mathcal{N}(\mu(n), c\mu(n)) \qquad (9)$$

LRHMM において，このような継続長を明示的に考慮するには，LRHMM の各状態に対して「次の状態に遷移するまでのフレーム数」を状態変数として付与することを考える．なお，このように，状態継続長を明示的に表すモデルは，Explicit-duration HMM と呼ばれ [19]，共通構造が既知である（楽譜情報が与えられている）場合のアライメント手法で，多く提案されている [9, 20]．

ここで，各状態に停留する時間が最大で L フレームであると仮定し，解釈別 LRHMM の状態空間を $N \times L$ の積空間に拡張すると，解釈別 LRHMM の状態系列を $S^{(d)}_{n,l}(t)$ と表すことができる．このとき，状態 $(n,1)$ から $(n+1,l)$ へ遷移する確率が $p(l|\mu(n),c)$ であり，状態 (n,l) から $(n,l-1)$ $(l>0)$ へ遷移する確率が 1 であり，それ以外の状態遷移確率を 0 とすれば，状態停留時間を明示的に扱える．$\mu(n)$ を，第二種の最尤推定で最適化すると，$\mu(n) = \frac{\sum_{d,t} l\langle S^{(d)}_{n,0}(t-1) S^{(d)}_{n,l}(t)\rangle}{\sum_{d,t}\langle S^{(d)}_{m,0}(t-1) S^{(d)}_{m,l}(t)\rangle}$ を得る．同様に c を推定することも可能であるが，予備実験によると，c は一定値（$c = 0.1$ 程度）に固定した方が，良好に動作することが示唆されている．

このようなモデルは，全ての音響信号が同一のテンポで演奏されることを仮定している．そのため，二つの演奏間のテンポが極端に違う場合に，性能が落ちる可能性がある．演奏速度の差を明示的に扱えるような，演奏長のモデルは，今後の検討課題である．

5. 評価実験

Chopin の Mazurka 9 曲（Op. 6-4, 17-4, 24-2, 30-2, 33-2, 41-2, 63-3, 67-1, 68-3）の各曲に対して，2 から 5 つの演奏録音（計 38 曲）を用意した．9 曲に対して，それぞれ (1) 二乗誤差最小化規準に基づく DTW，(2) 潜在共通構造として LRHMM を用いたもの（「LRHMM」），(3) 提案法（Proposed），(4) 提案法において，各楽曲に割り当てられる LRHMM を LRHSMM としてモデル化したもの（「Prop. (LRHSMM)」），(5) (4) において，潜在共通構造 HMM を NPHMM としてモデル化したもの

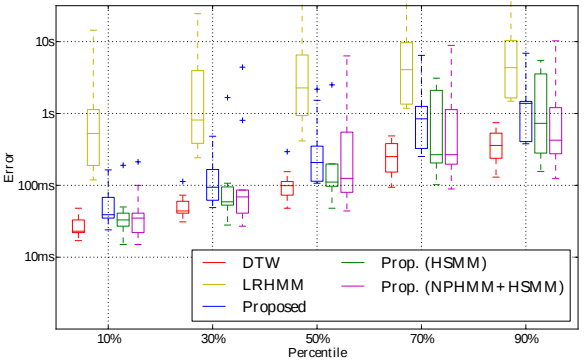

図 4 アライメント絶対誤差のパーセンタイル．

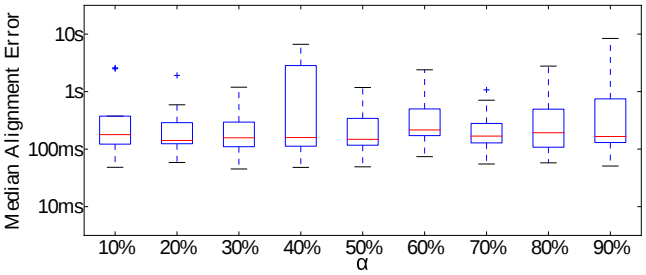

図 5 α に対するアライメント誤差中央値．

（「Prop. (NPHMM+LRHSMM)」），の 5 種類でアライメントを行った．

なお，$x(d,t)$ は，サンプリング周波数 44.1kHz，フレーム数 8192 サンプル，ホップ長 1764 サンプルで算出された Chroma ベクトル [21] とその時間方向の一次差分（Δchroma）の，計 24 次元の特徴量とした．

5.1 アライメント精度

アライメント精度を，Mazurka Project [1] で用意されている正解アライメントデータと比較した．

表 4 に，アライメント絶対誤差のパーセンタイルを表示する．この結果から，本手法は LRHMM よりも高い精度を持つことが分かる．このことから，生成モデルアプローチのアライメントにおいて，潜在共通構造を陽に考慮することの重要性が示唆された．

一方で，提案手法は DTW よりも精度が落ちる．これは，DTW が全体最適な手法であるのに対し，本手法は局所最適な手法であることに起因する．生成モデルのパラメータと，その状態系列を推定する過程において，本手法では誤推定が生じる可能性がある以上，適切な距離尺度が設定可能で，かつ生成モデルパラメータを推定する必要がない場合，DTW の方が適切であると言える．

5.2 楽曲進行の長さ N の設定に対する頑健性

本手法では，楽曲進行の長さ N を手動で設定する必要がある．この値に対する頑健さを検証するため，Proposed 手法に対して，N を $N = \alpha|T_{d=1}|$ とし，$\alpha = 0.1$ から $\alpha = 0.9$ の間で 0.1 刻みに走査したとき，それぞれの α におけるアライメント精度の中央値を求めた．**図 5** に結果を示す．この図から，α と推定精度は強く相関しないことが示唆されるため，推定精度はそれ以外の要因に強く影響されることが分かる．

表 1 オンセット検出率.

Method	Prec.	Rec.	F-meas.
LRHMM	43%	26%	31%
Proposed	72%	70%	70%
Prop. (LRHSMM)	75%	74%	74%
Prop. (PHMM+LRHSMM)	75%	74%	74%

5.3 オンセット検出精度

もし潜在共通構造 HMM の概念が妥当であるならば，潜在構造 HMM の状態が切り替わる地点では，瞬時に発生する音の特性が大きく異なることが想定される．そこで，潜在共通構造 HMM の状態変化地点をオンセットとしてみなしたときの，オンセット検出精度を評価する．そこで，Mazurka Project のオンセットデータ [1] を正解とした時，共通潜在構造の状態が変化する位置 [*2] を推定されたオンセット位置としたとき，それぞれ適合度，再現率と F 値を求めた．なお，正解オンセット位置から 0.1 秒以内に潜在構造 HMM の状態変化があった場合を正解とした．

表 1 に結果を示す．このことから，LRHMM ではオンセット検出精度が低く，状態の切り替わりと音の変化が対応づいていないことが示唆される．また，提案法では，各音響信号に割り当てられる LRHMM を LRHSMM にすることで，精度が向上する．これは，LRHSMM にすることにより，短時間でバースト状に発生する状態変化を抑えることができているためであると考えられる．また，Nonparametric HMM を用いても，性能差は大きく変化しないことも示唆された．

6. まとめ

本稿では，階層モデルに基づくオーディオ同士のアライメント手法を定式化した．本手法では，楽曲を表す構造である「潜在共通構造」を概念を導入することで，楽曲の基礎的な構成要素をコンパクトにモデル化した．また，楽曲を，「楽曲進行」という，潜在共通構造から出力された状態系列であるとみなし，アライメントを，楽曲進行のアライメントと置き換えた．すなわち，状態遷移順序の同一性を保ちながら，楽曲をコンパクトに表現することを両立した．このようなシンプルな階層モデルにより，生成モデルの学習に，より多くのデータを用いることができた．LRHMM と提案法を比較した実験によると，本研究により，生成モデルによるアプローチによる音響信号アライメントが，デファクト標準の DTW に大きく近づくことが示唆された．

今後の課題は，精度の向上，初期条件の違いに対して頑健さを確保することである．また，評価実験によれば，本手法により，確率生成モデルに基づく音響信号が，実用的な水準になったと考えられる．よって，今後は，より多くの，柔軟な生成モデルに，本手法を適用したい．

謝辞 Mazurka Project データを提供していただいた Craig S. Sapp 氏に感謝する．

[*2] 共通潜在構造では自分自身への遷移が許されるため，楽曲進行 n が切り替わる地点とは違うことに注意されたい

参考文献

[1] Sapp, C. S.: Comparative Analysis of Multiple Musical Performances, *ISMIR*, pp. 2–5 (2007).

[2] Miki, S. et al.: PEVI: Interface for retrieving and analyzing expressive musical performances with scape plots, *SMC*, pp. 748–753 (2013).

[3] Maezawa, A. et al.: Query-By-Conducting: An Interface to Retrieve Classical-music Interpretations by Real-time Tempo Input, *ISMIR*, pp. 477–482 (2010).

[4] Fremerey, C. et al.: A Demonstration of the SyncPlayer System, *ISMIR*, pp. 131–132 (2007).

[5] Nakamura, T. et al.: Acoustic Score Following to Musical Performance with Errors and Arbitrary Repeats and Skips for Automatic Accompaniment, *SMC*, pp. 200–304 (2013).

[6] 前澤　陽ほか：楽曲パート混合オーディオ同士の楽譜なしアライメント手法，情報処理学会音楽情報科学研究会, 2013-MUS-100 (2013).

[7] Miotto, R. et al.: Statistical Music Modeling Aimed at Identification and Alignment, *AdMiRE*, pp. 187–212 (2010).

[8] Duan, Z. et al.: Soundprism: An Online System for Score-Informed Source Separation of Music Audio, Vol. 5, No. 6, pp. 1205–1215 (2011).

[9] Raphael, C.: A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores, *ISMIR*, pp. 387–394 (2004).

[10] Dannenberg, R. B. et al.: Polyphonic Audio Matching for Score Following and Intelligent Audio Editors, *ICMC* (2003).

[11] Orio, N. et al.: Score Following : State of the Art and New Developments, *NIME*, pp. 36–41 (2003).

[12] Montecchio, N. et al.: A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Montecarlo inference techniques, *ICASSP*, pp. 193–196 (2011).

[13] Joder, C. et al.: Off-line refinement of audio-to-score alignment by observation template adaptation, *ICASSP*, pp. 206–210 (2013).

[14] Maezawa, A. et al.: Polyphonic Audio-to-score Alignment based on Bayesian Latent Harmonic Allocation Hidden Markov Model, *ICASSP*, pp. 185–188 (2011).

[15] Otsuka, T. et al.: Incremental Bayesian Audio-to-Score Alignment with Flexible Harmonic Structure Models, *ISMIR*, pp. 525–530 (2011).

[16] Paulus, J. et al.: State of the Art Report: Audio-Based Music Structure Analysis, *ISMIR*, pp. 625–636 (2010).

[17] Ferguson, T. S.: A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, Vol. 1, No. 2, pp. 209–230 (1973).

[18] Ding, N. et al.: Variational nonparametric Bayesian Hidden Markov Model, *ICASSP*, pp. 2098–2101 (2010).

[19] Yu, S. et al.: An efficient forward-backward algorithm for an explicit-duration hidden Markov model, *IEEE SPL*, Vol. 10, No. 1, pp. 11–14 (2003).

[20] Yamamoto, R. et al.: Robust on-line algorithm for real-time audio-to-score alignment based on a delayed decision and anticipation framework, *ICASSP*, pp. 191–195 (2013).

[21] Fujishima, T.: Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music, *ICMC*, pp. 464–467 (1999).