

# 同質性・反復性・規則性を考慮した 階層隠れセミマルコフモデルに基づく統計的音楽構造解析

柴田 剛<sup>1,a)</sup> 錦見 亮<sup>1,b)</sup> 中村 栄太<sup>1,2,c)</sup> 吉井 和佳<sup>1,d)</sup>

受付日 2019年7月5日, 再受付日 2019年7月5日,  
採録日 2019年7月5日

**概要:** 本稿では, 音楽音響信号を音楽的に意味のあるひとまとまりの区間 (セクション) に分割し, それらをいくつかのクラスに分類する音楽構造解析手法について述べる. 我々は, 音楽構造を決定する三つの基本的側面, 即ち各セクション内における音色の同質性, 同じクラスのセクションにおけるコード進行の反復性, およびセクション長の規則性に着目し, これらを確率的な枠組みで統一的に取り扱うための階層隠れセミマルコフモデルを提案する. 本モデルは, セクション系列とコード系列に対応する二階層の潜在変数系列を持ち, 音色特徴量 (メル周波数ケプストラム係数) とコード特徴量 (クロマベクトル) を観測変数系列として出力する. まず, 上位のセクション系列は, 各セクションの継続時間長を考慮したセミマルコフモデルに従うと仮定し, 音色の同質性を担保するため, セクションクラスごとに音色特徴量の出力分布を仮定する. 一方, 下位のコード系列は, 同じクラスのセクションでは同じ順序でコード進行が反復されるように, セクション条件付き Left-to-Right 型マルコフモデルに従うと仮定する. 各パラメータに共役事前分布を導入してベイズモデルを構成することにより, セクション数とコード数を過剰に設定しても, 観測データに合わせて適切な個数のセクションとコードからなる潜在変数系列を推定できる. 実験により, 同質性と規則性の統合による性能向上を確認した. また, 提案法による音楽構造解析結果は正解データと類似する統計的性質を持ち, 分割・分類精度において代表的な既存手法より優れていることを確認した.

**キーワード:** 音楽構造解析, 隠れセミマルコフモデル, 教師なし学習, 統計的音楽信号処理, ベイズ推論

## Statistical Method for Music Structure Analysis Based on a Hierarchical HSMM

GO SHIBATA<sup>1,a)</sup> RYO NISHIKIMI<sup>1,b)</sup> EITA NAKAMURA<sup>1,2,c)</sup> KAZUYOSHI YOSHII<sup>1,d)</sup>

Received: July 5, 2019, Revised: July 5, 2019,  
Accepted: July 5, 2019

**Abstract:** This paper describes a music structure analysis method that splits music audio signals into meaningful segments (musical sections) and clusters them. Focusing on three fundamental aspects that characterize musical structures, *homogeneity* of timbre within each section, *repetitiveness* of chord progression in sections of the same class, and *regularity* of durations of sections, we propose a hierarchical hidden semi-Markov model (HSMM) that can deal with these aspects in a unified probabilistic framework. This model has two sequences of latent states corresponding to a sequence of sections and that of chords. The timbral features (mel-frequency cepstrum coefficients) and chord features (chroma vectors) are emitted as observed variables. The higher-level sequence of sections is assumed to follow a semi-Markov model that explicitly represents the duration of each section. The emission distributions of timbral features are assigned to individual section classes to guarantee the homogeneity of timbre. The lower-level sequence of chords is assumed to follow a section-conditioned left-to-right Markov model. This model represents the repetition of chord progressions in sections of the same class. We formulate a Bayesian model by putting conjugate prior distributions. The sequences of latent states with appropriate effective numbers of sections and chords can be estimated even if too many sections and chords are assumed. Evaluation experiments showed that the joint modeling of homogeneity and regularity improved the performance. In addition, the proposed method can yield analysis results with similar statistical properties as the ground truth data and has higher accuracy than conventional methods in segmentation and clustering.

**Keywords:** music structure analysis, hidden semi-Markov model, unsupervised learning, statistical musical signal processing, Bayesian inference

# 1. はじめに

音楽音響信号からセクション（ポピュラー音楽の A メロ, B メロ, サビなど）と呼ばれる意味のあるセグメントの検出をする音楽構造解析 [1] は、音楽情報検索 (Music Information Retrieval; MIR) の基礎技術であり、長年研究されているトピックである。一般に、音楽構造解析は音楽音響信号をセクションに分割する「セグメンテーションステップ」[2-9]、各セクションをいくつかのクラスに分類する「クラスタリングステップ」[10-18]、そして各クラスに A メロ, サビのような具体的なラベルを付ける「ラベリングステップ」[19-21] を含む。本稿では、ポピュラー音楽のセグメンテーションとクラスタリングを取り扱う。

従来、ポピュラー音楽のセクションはセクション内の「同質性」、セクション間の「反復性」と「新規性」の三つの側面を持つとされてきた [1]。具体的には、同質性は音響特性（例えば、メル周波数ケプストラム係数 (Mel-Frequency Cepstrum Coefficient; MFCC) などの音色特徴量）がセクション内で一貫していることを意味する。反復性は、セクションのある音楽的要素の系列（例えば、クロマベクトルやコード進行）が同じクラスのセクションで繰り返されることを意味する。新規性は、音楽特性がセクションの境界で急激に変化することを意味する。さらに、ポピュラー音楽ではセクション長はその多くが 4 または 8 小節であり、いくつかの研究ではそのような「規則性」に注目している [8-10]。しかし、第 2 章で説明するように、音楽構造解析に関する研究は上記の側面のうち一つだけに注目するか、または複数の側面を別々に扱うものがほとんどである。これら四つの側面を同時に捉える計算モデルの構築が音楽構造解析における中心的な課題である。

本研究では、確率的な枠組みでセクションの同質性・反復性・規則性を同時に扱う統計的音楽構造解析手法を提案する (図 1)。具体的には、セクション、コード進行、および音楽音響信号（音色特徴量とクロマベクトル）の階層的生成過程を表す、階層隠れセミマルコフモデル (Hierarchical Hidden Semi-Markov Model; HHSMM) と呼ばれる統一的確率モデルを定式化する。このモデルは二階層の潜在状態系列を持つ。上位の系列はセクション系列を表し、継続長の規則性を明示的に表現するセミマルコフモデルで記述される。下位の系列はコード系列を表し、セクションに条件付いた Left-to-Right 型のマルコフモデルで記述される。セクション内の音色特徴量の同質性を表すために、MFCC

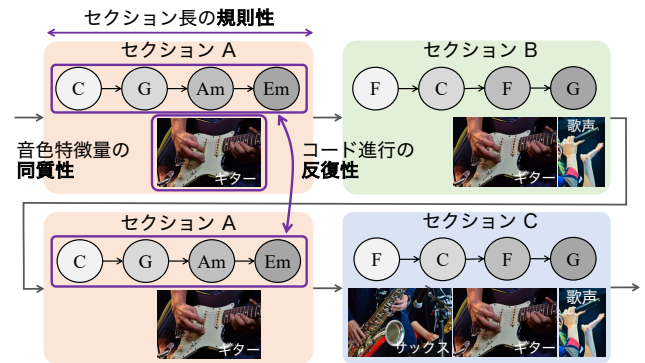


図 1 音色特徴量の同質性、コード進行の反復性、およびセクション長の規則性に基づく音楽構造解析。

はセクションに対応する上位状態から生成されると仮定する。コード進行のセクション間の反復性を表すために、セクションのクロマベクトルは下位状態から系列的に生成されると仮定する。観測データとして音楽音響信号が与えられた時、ギブスサンプリングとビタビ学習を用いることでモデル全体を教師なしで学習できる。この際、ベイズ推論に基づくスパース学習の効果により、最適なセクション数が自動的に推定される。

本研究の主な貢献は、セクションの同質性・反復性・規則性を統一的に扱える生成モデルに基づく音楽構造解析手法の提案である。このアプローチは、セクションのアノテーションを用いた教師あり学習に基づく深層識別モデル [6-8] と異なり、教師なし学習を実行できる利点を持つ。これら二つのアプローチは相互補完的な関係にあるため、本研究の結果は識別モデルと生成モデルの深層ベイズ統合である変分自己符号化の枠組み [22] (音響信号からセクション系列, 反対にセクション系列から音響信号のモデル化) によるさらなる改善の可能性につながる。もう一つの重要な貢献は、各種手法の推定結果の統計的特徴を詳細に調査したことである。本稿では、提案法により推定されたセクション長、セクションクラス数、およびセクション境界の拍節位置の分布が、従来法から得られる分布よりも正解データと近いものとして得られることを示す。なお、本稿は、国際会議における我々の報告 [23] をベースに、推定結果に対する詳細なエラー解析を加えたものである。

## 2. 関連研究

音楽構造解析に対する最も標準的なアプローチは、クロマベクトルや MFCC などの音響特徴量の自己類似度行列 (Self-Similarity Matrix; SSM) を用いるものである。この行列の各要素は、二つの時間フレーム間の音響的類似性を表す (図 2)。SSM では、同質性、反復性、新規性、および規則性は、それぞれブロック対角構造、対角線に平行な短い縞、格子パターン、および格子間隔の規則性として現れる。これら四つの側面は音楽構造解析におけるセグメンテーションおよびクラスタリングタスクに使用されてきた。

<sup>1</sup> 京都大学 大学院情報学研究所  
Graduate School of Informatics, Kyoto University  
<sup>2</sup> 京都大学 白眉センター  
The Hakubi Center for Advanced Research, Kyoto University  
a) gshibata@sap.ist.i.kyoto-u.ac.jp  
b) nishikimi@sap.ist.i.kyoto-u.ac.jp  
c) enakamura@sap.ist.i.kyoto-u.ac.jp  
d) yoshii@kuis.kyoto-u.ac.jp

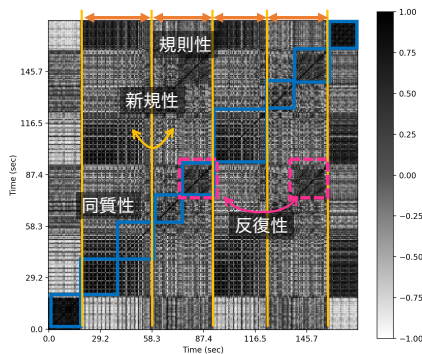


図 2 MFCC による自己類似度行列 (Self-Similarity Matrix; SSM) (RWC-MDB-P-2001 No. 25 の一部).

## 2.1 セグメンテーション

Foot 2] は新規性に着目し、チェッカーボードカーネルを SSM の対角要素に沿って畳み込んで得られる時変新規性曲線からピークを検出する方法を提案した. Jensen [3] は、同質性と新規性に基づくコスト関数の最小化によるセクション境界推定法を提案した. Goto [19] は、反復構造が縞ではなく垂直線として現れる様に工夫されたラグ SSM およびこれに基づく新規性曲線の計算法を提案した. Serra [4] は局所的な変化と大域的な特徴の両方を捉える新たな新規性曲線を提案した. Peeters ら [5] は、これらの二つの手法 [4, 19] を統合し、セグメンテーションの改良を行った. Ullrich ら [6] は、畳み込みニューラルネットワーク (CNN) に基づく教師あり学習手法の先駆けとなった. この手法は、粗いレベルと細かいレベルの両方の境界アンテーションを取り扱う手法に拡張されている [7]. Smith ら [24] は、複数の音楽構造解析結果から最適なものを選択する後処理に規則性を用いることは性能向上に有効ではなく、解析時に規則性を考慮することが重要である可能性に言及した. Sargent ら [9] は、一曲中のセクションは同じ継続長を持ちやすいという規則性を捉えることの有効性を指摘した. Maezawa [8] は、同質性・反復性・新規性・規則性に基づくコスト関数と LSTM (長短期記憶) ネットワークに基づく手法を開発した. これら関連研究の知見を参考にして、本研究では、同質性・反復性・規則性を同時に考慮した生成モデルに基づく手法を提案する.

## 2.2 クラスタリング

Cooper ら [12] は、セグメンテーション [2] とセクション内部およびセクション間の統計的特性に基づくクラスタリングを段階的に行った. Goodwin ら [13] は、動的計画法を用いて SSM の非対角成分の中に縞の構造を効率的に検出する試みを行った. 反復性と同質性を同時に扱うため、Grohganz ら [14] は SSM の持つ対角線に平行な縞を固有値分解を用いてブロック対角構造に変換することで、同質性に基づく手法が反復性にも適用できることを示した. Nieto ら [15] は、非負値行列分解に凸結合の制約を加

え、セクション境界の検出とセクションのクラスタリングを行った. McFee ら [16] は、反復構造をグラフを用いて記述し、グラフ分解のためのスペクトル・クラスタリングを用いる方法を提案した.

セグメンテーションとクラスタリングを同時に行う生成モデルに基づく統計的手法に関する研究も行われている. Aucouturier ら [11] は、標準的な HMM に基づく手法を調べた. Levy ら [25] は、隠れセミマルコフモデルに基づくセクション長の規則性を考慮する手法を提案した. Ren ら [17] は、セクションの個数を推定できる HMM のノンパラメトリックベイズ拡張を提案した. Barrington ら [18] も、自動的にモデルの複雑度を制御できるスイッチング線形動的システムのノンパラメトリックベイズ拡張を提案している. これらの方法は主に同質性と規則性に着目している一方で、提案法は反復性も同時に考慮でき、継続長や拍節位置などのセクションに関する統計的特徴に関する事前知識を取り込んだベイズ推論を行える利点がある.

## 3. 提案法

本章では、提案する統計的音楽構造解析手法を説明する. 提案法は、単一楽曲生成モデルの教師なしベイズ学習に基づいている. セクションの継続時間長の事前分布を定義する際に、解析対象の楽曲を含まない楽曲群 (正解のセクション系列) から予め学習しておいた経験分布を用いる. それ以外のハイパーパラメータは、事前知識を用いず手動で設定する. 対象楽曲に対してパラメータの事後分布推定を行うことで、事後確率が最大となるセクション系列を得る.

### 3.1 問題設定

我々が取り組む問題を以下のように定式化する.

入力: 音楽音響信号から得られた、ビート単位のクロマベクトル系列  $\mathbf{X}^c = \mathbf{x}_{1:B}^c \in \mathbb{R}^{B \times 12}$  と MFCC 系列  $\mathbf{X}^m = \mathbf{x}_{1:B}^m \in \mathbb{R}^{B \times 12}$

出力: セクションの境界とクラス

ここで、 $B$  は四分音符単位のビート数、添え字  $\circ_{a:b}$  は系列  $(\circ_a, \dots, \circ_b)$  を表す. 本手法では、オクターブ内の 12 種類のピッチクラスに対応する 12 次元クロマベクトルと、音色特徴量として 12 次元 MFCC を使用する.

### 3.2 モデル定式化

図 3 に示すように、提案モデルを二階層のマルコフ連鎖と音響モデルで構成する. 上位のマルコフ連鎖はセクションレベルの構造 (セクションクラスと継続長) を表現し、下位のマルコフ連鎖は各セクションの内部構造 (コード進行) を表現する. 音響モデルは、これらの潜在状態と観測された音楽特徴量 (クロマベクトルと MFCC) との関係性を記述する.

### 3.2.1 上位マルコフ連鎖

上位のマルコフ連鎖は全遷移型のセミマルコフモデルであり、セクション系列  $\mathbf{Z} = z_{1:T}$  ( $z_\tau \in \{1, \dots, N_Z\}$ ) と継続長系列  $\mathbf{D} = d_{1:T}$  ( $d_\tau \in \{1, \dots, N_D\}$ ) を生成する。ここで、 $T$  はセクションの数、 $N_Z$  は取りうるセクションクラスの種類数、 $N_D$  はセクションの最大継続長を表す。セクション系列と継続長系列の生成過程は以下の通りである。

$$p(z_1, d_1) = \rho_{z_1} \psi_{d_1} \quad (1)$$

$$p(z_\tau, d_\tau | z_{\tau-1}, d_{\tau-1}) = \pi_{z_{\tau-1} z_\tau} \psi_{d_\tau} \quad (2)$$

ここで、 $\rho_z$  と  $\pi_{zz'}$  はセクション系列の初期確率と遷移確率であり、 $\psi_d$  は継続長確率である。

### 3.2.2 下位マルコフ連鎖

下位のマルコフ連鎖は状態数  $N_K$  の Left-to-Right 型マルコフモデルであり、各セクションの内部構造を表現する。各状態はコードを表し、状態系列はコード進行を表す。各セクションがこのようなマルコフ連鎖を持ち、対応するセクションの開始時刻から継続時間が経過するまでビート単位で状態遷移を続ける。状態系列  $\mathbf{K}_\tau = k_{\tau,1:d_\tau}$  ( $k_{\tau,t} \in \{1, \dots, N_K\}$ ) の生成過程は以下の通りである。

$$p(k_{\tau,t} | z_\tau, k_{\tau,t-1}) = \phi_{k_{\tau,t-1} k_{\tau,t}}^{(z_\tau)} \quad (3)$$

ここで、 $z_\tau$  と  $d_\tau$  は対応するセクションのクラスと継続長であり、 $\phi_{kk'}^{(z)}$  は状態  $k$  から状態  $k'$  への遷移確率である。

この Left-to-Right 型マルコフモデルは、初期状態が  $k_{\tau,1} = 1$  で、かつ、 $t_1 < t_2$  ならば  $k_{\tau,t_1} \leq k_{\tau,t_2}$  を満たす。このように、同じクラスのセクション同士では類似するコード進行を持つという制約を与えることで反復性を表現する。また、状態遷移の最大幅を表すハイパーパラメータ  $\sigma$  を導入し、状態  $k$  から状態  $k + \sigma$  までの遷移は許容するが、それより大きな状態遷移を禁止する。すなわち、 $k' > k + \sigma$  の時  $\phi_{kk'}^{(z)} = 0$  である。これにより、反復構造における揺らぎを表現する。3.2.4 項と 3.3.2 項で後述する通り、本モデルでは遷移確率  $\phi_{kk'}^{(z)}$  の値を直接は与えず、事前分布をおいた上で入力データに応じて学習する。以降、 $\mathbf{K}_{1:T}$  を  $\mathbf{K}$  と表す。

### 3.2.3 音響モデル

音響モデルは、セクションのクラス  $\mathbf{Z}$  と内部状態  $\mathbf{K}$  に条件付けられた出力確率を用いて、クロマベクトル  $\mathbf{x}_b^c \in \mathbb{R}^{12}$  と MFCC  $\mathbf{x}_b^m \in \mathbb{R}^{12}$  の生成過程を表現する。コード進行の系列構造を表現するため、クロマベクトルの出力確率  $\chi_{z,k}^c$  は  $\mathbf{Z}$  と  $\mathbf{K}$  の両方に依存すると仮定する。また、各セクションにおける音色特徴量の同質性を捉えるため、MFCC の出力確率  $\chi_z^m$  は  $\mathbf{Z}$  にのみ依存すると仮定する。

$$p(\mathbf{x}_b^c, \mathbf{x}_b^m) = \chi_{z_b, k_b}^c(\mathbf{x}_b^c) \chi_{z_b}^m(\mathbf{x}_b^m) \quad (4)$$

ここで、 $z_b$  と  $k_b$  はそれぞれビート  $b$  におけるセクション

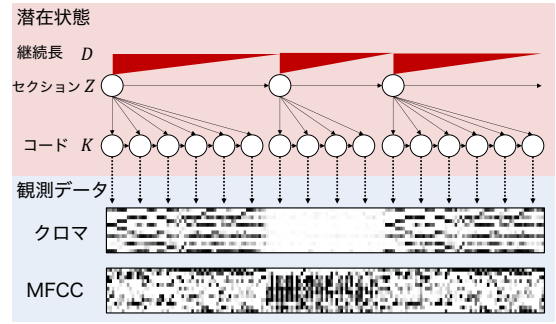


図 3 提案モデルにおける音楽音響信号の生成過程

のクラスと内部状態を表す。出力確率は多変量正規分布に従うことを仮定する。

$$\chi_{z,k}^c(\mathbf{x}^c) = \mathcal{N}(\mathbf{x}^c | \boldsymbol{\mu}_{z,k}^c, (\boldsymbol{\Lambda}_{z,k}^c)^{-1}) \quad (5)$$

$$\chi_z^m(\mathbf{x}^m) = \mathcal{N}(\mathbf{x}^m | \boldsymbol{\mu}_z^m, (\boldsymbol{\Lambda}_z^m)^{-1}) \quad (6)$$

ここで、 $\boldsymbol{\mu}_{z,k}^c$  と  $\boldsymbol{\Lambda}_{z,k}^c$  はクロマベクトルの平均と精度行列、 $\boldsymbol{\mu}_z^m$  と  $\boldsymbol{\Lambda}_z^m$  は MFCC の平均と精度行列である。

### 3.2.4 事前分布

共役事前分布を置くことでベイズ HHSMM を定式化する。離散分布に対してはディリクレ事前分布を置く。

$$\boldsymbol{\rho} \sim \text{Dirichlet}(\mathbf{a}^\rho) \quad (7)$$

$$\boldsymbol{\psi} \sim \text{Dirichlet}(\mathbf{a}^\psi) \quad (8)$$

$$\boldsymbol{\pi}_z \sim \text{Dirichlet}(\mathbf{a}^\pi) \quad (9)$$

$$\phi_k^{(z)} \sim \text{Dirichlet}(\mathbf{a}^\phi) \quad (10)$$

ここで、 $\boldsymbol{\rho} = \rho_{1:N_Z}$ 、 $\boldsymbol{\psi} = \psi_{1:N_D}$ 、 $\boldsymbol{\pi}_z = \pi_{z(1:N_Z)}$ 、 $\phi_k^{(z)} = \phi_{k(1:N_K)}^{(z)}$  であり、 $\mathbf{a}^\rho$ 、 $\mathbf{a}^\psi$ 、 $\mathbf{a}^\pi$ 、および  $\mathbf{a}^\phi$  はハイパーパラメータである。これらのパラメータの値が小さい場合、クラス間の遷移確率はまばらになる。これにより、モデルは小さい音響的変動があっても反復構造を捉えられ、不必要なセクションクラスを取り除ける。

ポピュラー音楽では、セクション長は 4 小節の整数倍になる傾向があるので (図 4)、そうした統計的な傾向を事前分布に組み込むことができる。具体的には、 $\mathbf{a}^\psi$  としてセクション長の経験分布  $\mathbf{a}_{\text{emp}}^\psi$  を定数倍したものをを用いる。セクションクラスの構造は個別の楽曲によって大きく異なるため、初期確率および遷移確率に対しては一様なディリクレ事前分布を仮定する。

最後に、多変量正規分布に対してはガウス・ウィンシャート事前分布を置く。

$$\boldsymbol{\mu}_{z,k}^c, \boldsymbol{\Lambda}_{z,k}^c \sim \mathcal{N}(\boldsymbol{\mu}_{z,k}^c | \mathbf{m}_0^c, (\beta_0^c \boldsymbol{\Lambda}_{z,k}^c)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_{z,k}^c | \mathbf{W}_0^c, \nu_0^c)$$

$$\boldsymbol{\mu}_z^m, \boldsymbol{\Lambda}_z^m \sim \mathcal{N}(\boldsymbol{\mu}_z^m | \mathbf{m}_0^m, (\beta_0^m \boldsymbol{\Lambda}_z^m)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_z^m | \mathbf{W}_0^m, \nu_0^m)$$

ここで、 $\mathbf{m}_0^c$ 、 $\beta_0^c$ 、 $\mathbf{W}_0^c$ 、 $\nu_0^c$ 、 $\mathbf{m}_0^m$ 、 $\beta_0^m$ 、 $\mathbf{W}_0^m$  および  $\nu_0^m$  はハイパーパラメータである。



### 3.3 ベイズ学習

我々の目的は事後分布  $p(\mathbf{Z}, \mathbf{D}, \mathbf{K}, \Theta | \mathbf{X}^c, \mathbf{X}^m)$  の計算である。ここで、 $\Theta = \{\rho, \psi, \pi, \phi, \mu, \Lambda\}$  である。この事後分布は解析的に計算できないため、ギブスサンプリング法を用いる。分布  $p(\mathbf{Z}, \mathbf{D}, \mathbf{K} | \Theta, \mathbf{X}^c, \mathbf{X}^m)$  から潜在変数  $\mathbf{Z}$ ,  $\mathbf{D}$ ,  $\mathbf{K}$  をサンプルしたのち、分布  $p(\Theta | \mathbf{Z}, \mathbf{D}, \mathbf{K}, \mathbf{X}^c, \mathbf{X}^m)$  からモデルパラメータ  $\Theta$  をサンプルする。この処理を反復することで、真の事後分布からのサンプルを得る。

#### 3.3.1 潜在変数のサンプリング

上位と下位の潜在変数  $\mathbf{Z}$ ,  $\mathbf{D}$ ,  $\mathbf{K}$  のサンプリングにはフォワードフィルタリング・バックワードサンプリング法を用いる。ビート  $b - d_b + 1$  から始まり、ビート  $b$  で終わるセクションのクラスと継続長を表す変数  $z_b$  と  $d_b$  を導入する。また、このセクションの周辺出力確率を以下のようにおく。

$$\begin{aligned} & \omega_{z_b}(\mathbf{x}_{b-d_b+1:b}^c, \mathbf{x}_{b-d_b+1:b}^m) \\ &= \sum_{k_{b-d_b+1:b} \in \{1, \dots, N_K\}^{d_b}} \prod_{t=1}^{d_b-1} \phi_{k_{b-d_b+t} k_{b-d_b+t+1}}^{(z_b)} \\ & \quad \cdot \prod_{t=1}^{d_b} \chi_{z_b, k_{b-d_b+t}}^c(\mathbf{x}_{b-d_b+t}^c) \chi_{z_b}^m(\mathbf{x}_{b-d_b+t}^m) \quad (11) \end{aligned}$$

この確率は下位のマルコフ連鎖に対してフォワードアルゴリズムを用いることで計算できる。

上位モデルのフォワードフィルタリングステップでは、変数  $\alpha_b(z_b, d_b) = p(z_b, d_b, \mathbf{x}_{1:b}^c, \mathbf{x}_{1:b}^m)$  の初期化と更新を行う。

$$\alpha_b(z_b, d_b = b) = \rho_{z_b} \psi_{d_b} \omega_{z_b}(\mathbf{x}_{1:b}^c, \mathbf{x}_{1:b}^m) \quad (12)$$

$$\alpha_b(z_b, d_b) \quad (13)$$

$$= \sum_{z', d'} \alpha_{b-d_b}(z', d') \pi_{z' z_b} \psi_{d_b} \omega_{z_b}(\mathbf{x}_{b-d_b+1:b}^c, \mathbf{x}_{b-d_b+1:b}^m)$$

バックワードサンプリングステップでは、潜在変数  $\mathbf{Z}$  と  $\mathbf{D}$  を後ろから順にサンプリングする。具体的には、変数  $z_b$  と  $d_b$  がサンプル済みであるとき、ビート  $b' = b - d_b$  における変数  $z_{b'}$  と  $d_{b'}$  をサンプルする。

$$p(z_B, d_B | \mathbf{X}^c, \mathbf{X}^m) \propto \alpha_B(z_B, d_B) \quad (14)$$

$$p(z_{b'}, d_{b'} | z_{b:B}, d_{b:B}, \mathbf{X}^c, \mathbf{X}^m) \propto \alpha_{b'}(z_{b'}, d_{b'}) \pi_{z_{b'} z_b} \quad (15)$$

次に、潜在変数  $\mathbf{K}$  をサンプル済みの  $\mathbf{Z}$  と  $\mathbf{D}$  を用いてサンプリングする。各変数集合  $\mathbf{K}_\tau$  は、クラス  $z_\tau$  の下位モデルに対してフォワードフィルタリング・バックワードサンプリング法を用いることでサンプリングする。ここで、セクション境界に対して相対的に定められるビート添字  $t \in \{1, \dots, d_\tau\}$  を用いる。フォワードフィルタリングステップでは、確率  $\zeta_{\tau, k_\tau, t}$  を再帰的に計算する。

$$\begin{aligned} \zeta_{\tau, k_\tau, 1} &= p(k_\tau, 1, \mathbf{x}_1^c, \mathbf{x}_1^m | z_\tau, d_\tau) \\ &= \delta_{k_\tau, 1} \chi_{z_\tau, 1}^c(\mathbf{x}_1^c) \chi_{z_\tau}^m(\mathbf{x}_1^m) \quad (16) \end{aligned}$$

$$\begin{aligned} \zeta_{\tau, k_\tau, t} &= p(k_\tau, t, \mathbf{x}_{1:t}^c, \mathbf{x}_{1:t}^m | z_\tau, d_\tau) \\ &= \left( \sum_{k_\tau, t-1} \zeta_{\tau, k_\tau, t-1} \phi_{k_\tau, t-1 k_\tau, t}^{(z_\tau)} \right) \chi_{z_\tau, k_\tau, t}^c(\mathbf{x}_t^c) \chi_{z_\tau}^m(\mathbf{x}_t^m) \quad (17) \end{aligned}$$

バックワードサンプリングでは、潜在変数  $\mathbf{K}_\tau$  を後ろから順にサンプリングする。

$$p(k_\tau, d_\tau | z_\tau, d_\tau, \mathbf{x}_{1:d_\tau}^c, \mathbf{x}_{1:d_\tau}^m) \propto \zeta_{\tau, k_\tau, d_\tau} \quad (18)$$

$$p(k_\tau, t | z_\tau, d_\tau, k_\tau, t+1:d_\tau, \mathbf{x}_{1:d_\tau}^c, \mathbf{x}_{1:d_\tau}^m) \propto \zeta_{\tau, k_\tau, t} \phi_{k_\tau, t k_\tau, t+1}^{(z_\tau)} \quad (19)$$

#### 3.3.2 モデルパラメータのサンプリング

本研究では、ギブスサンプリング法を用いる。

$$\rho \sim \text{Dirichlet}(\mathbf{a}^\rho + \mathbf{b}^\rho) \quad (20)$$

$$\pi_z \sim \text{Dirichlet}(\mathbf{a}^\pi + \mathbf{b}^{\pi_z}) \quad (21)$$

$$\psi \sim \text{Dirichlet}(\mathbf{a}^\psi + \mathbf{b}^\psi) \quad (22)$$

$$\phi_k^{(z)} \sim \text{Dirichlet}(\mathbf{a}^\phi + \mathbf{b}^{\phi_k^{(z)}}) \quad (23)$$

$$\Lambda_{z,k}^c \sim \mathcal{W}(\mathbf{W}_{z,k}^c, \nu_{z,k}^c) \quad (24)$$

$$\mu_{z,k}^c | \Lambda_{z,k}^c \sim \mathcal{N}(\mathbf{m}_{z,k}^c, (\beta_{z,k}^c \Lambda_{z,k}^c)^{-1}) \quad (25)$$

$$\Lambda_z^m \sim \mathcal{W}(\mathbf{W}_z^m, \nu_z^m) \quad (26)$$

$$\mu_z^m | \Lambda_z^m \sim \mathcal{N}(\mathbf{m}_z^m, (\beta_z^m \Lambda_z^m)^{-1}) \quad (27)$$

ここで、 $\mathbf{b}^\rho \in \mathbb{R}^{N_Z}$ ,  $\mathbf{b}^{\pi_z} \in \mathbb{R}^{N_Z}$ ,  $\mathbf{b}^\psi \in \mathbb{R}^{N_D}$ ,  $\mathbf{b}^{\phi_k^{(z)}} \in \mathbb{R}^{N_K}$  である。 $b_z^1$  は  $z = z_1$  の時に 1, そうでなければ 0 となる。また、 $b_z^{\pi_z}$  は状態  $z$  から状態  $z'$  への遷移回数,  $b_d^z$  は継続長  $d$  がサンプルされた回数,  $b_{k'}^{\phi_k^{(z)}}$  はセクション  $z$  の下位モデルにおける状態  $k$  から状態  $k'$  への遷移回数を表す。パラメータ  $\mathbf{m}_{z,k}^c$ ,  $\beta_{z,k}^c$ ,  $\mathbf{W}_{z,k}^c$ ,  $\nu_{z,k}^c$  は次式で計算できる。

$$\beta_{z,k}^c = \beta_0^c + N_{z,k}, \quad \nu_{z,k}^c = \nu_0^c + N_{z,k} \quad (28)$$

$$\mathbf{m}_{z,k}^c = \frac{1}{\beta_{z,k}^c} (\beta_0^c \mathbf{m}_0^c + N_{z,k} \bar{\mathbf{x}}_{z,k}^c) \quad (29)$$

$$\begin{aligned} (\mathbf{W}_{z,k}^c)^{-1} &= (\mathbf{W}_0^c)^{-1} + N_{z,k} \mathbf{S}_{z,k}^c \\ &+ \frac{\beta_0^c N_{z,k}}{\beta_0^c + N_{z,k}} (\bar{\mathbf{x}}_{z,k}^c - \mathbf{m}_0^c)(\bar{\mathbf{x}}_{z,k}^c - \mathbf{m}_0^c)^\top \quad (30) \end{aligned}$$

ここで、 $N_{z,k}$ ,  $\bar{\mathbf{x}}_{z,k}^c$ ,  $\mathbf{S}_{z,k}^c$  をそれぞれ以下のように定義する。

$$N_{z,k} = \sum_{b=1}^B \delta_{z_b, z} \delta_{k_b, k} \quad (31)$$

$$\bar{\mathbf{x}}_{z,k}^c = \frac{1}{N_{z,k}} \sum_{b=1}^B \delta_{z_b, z} \delta_{k_b, k} \mathbf{x}_b^c \quad (32)$$

$$\mathbf{S}_{z,k}^c = \frac{1}{N_{z,k}} \sum_{b=1}^B \delta_{z_b, z} \delta_{k_b, k} (\mathbf{x}_b^c - \bar{\mathbf{x}}_{z,k}^c)(\mathbf{x}_b^c - \bar{\mathbf{x}}_{z,k}^c)^\top \quad (33)$$

パラメータ  $\mathbf{m}_z^m$ ,  $\beta_z^m$ ,  $\mathbf{W}_z^m$ ,  $\nu_z^m$  も同様に計算される。

### 3.3.3 モデルの改良

学習を促進するために二つの改良を導入する。まず、パラメータ推定の最終ステップでビタビ学習 [26] を適用する。ギブスサンプリングで推定されるパラメータは必ずしも事後分布の局所最適解ではない。そこで、潜在変数の推定にはフォワードフィルタリング・バックワードサンプリングアルゴリズムの代わりにビタビアルゴリズムを適用し、潜在変数に関する事後確率を最大化する。また、モデルパラメータの更新には事後分布からのサンプルの代わりに事後分布の期待値を用いる。

次に、継続長確率に対して重み係数  $w_{dur} (\geq 1)$  を導入して、その確率の影響を強める。具体的には、式 (12) と (13) において確率係数  $\psi_{d_b}$  を  $(\psi_{d_b})^{w_{dur}}$  で置き換える。前述のビタビ学習ステップや 3.4 節で述べるセクション推定ステップにおいても、同様の置き換えを行う。重み係数を大きくすることで、より頻出する継続長に重点を置く効果が得られる。

### 3.4 セクション推定

モデルパラメータを学習したのち、潜在変数（セクション）を最大事後確率（maximum a posteriori; MAP）推定によって求める。具体的には、潜在変数  $\mathbf{Z}$  と  $\mathbf{D}$  に関して事後確率  $p(\mathbf{Z}, \mathbf{D} | \Theta, \mathbf{X}^c, \mathbf{X}^m)$  を最大化する。これは、下位状態  $\mathbf{K}$  を積分消去し、上位モデルに対して隠れセミマルコフモデルのビタビアルゴリズム [27] を適用することで求められる。

## 4. 評価実験

本章では、提案手法の評価実験について述べる。

### 4.1 実験条件

評価には RWC 音楽データベース [28] とその構造アノテーションデータ [29] を用いた。簡単のため、100 曲あるデータのうち曲全体が 4/4 拍子である 85 曲を用いた。提案法の入力に用いる特徴量として、クロマベクトルには深層特徴量抽出 [30] の結果を用い、MFCC には librosa ライブラリ [31] の出力結果を用いた。ビート単位の特徴量を得るため、madmom ライブラリ [32] により得られたビート情報を用いて平均をとった。セクション長の経験分布  $\mathbf{a}_{emp}^{\psi}$  は 85 曲に対する leave-one-out 交差検証によって学習し、解析対象の楽曲はこの学習には用いないものとした。パラメータ推定では、ギブスサンプリングを 15 回、ビタビ学習を 3 回反復した。これには標準的な CPU において入力信号長の約 5 倍の時間を要した。提案法のハイパーパラメータは以下のようにした。  $N_Z = 12$ ,  $N_D = 40$ ,  $N_K = 16$ ,  $\mathbf{a}^{\rho} = 0.1 \cdot \mathbb{I}$ ,  $\mathbf{a}^{\pi} = \mathbb{I}$ ,  $\mathbf{a}^{\psi} = 50 \cdot \mathbf{a}_{emp}^{\psi}$ ,  $\mathbf{a}^{\phi} = \mathbb{I}$ ,  $\mathbf{m}_0^c = \mathbb{E}[\mathbf{X}^c]$ ,  $\beta_0^c = 8$ ,  $\mathbf{W}_0^c = (\nu_0^c \text{cov}[\mathbf{X}^c])^{-1}$ ,  $\nu_0^c = 96$ ,  $\mathbf{m}_0^m = \mathbb{E}[\mathbf{X}^m]$ ,  $\beta_0^m = 4$ , and  $\mathbf{W}_0^m = (\nu_0^m \text{cov}[\mathbf{X}^m])^{-1}$ ,  $\nu_0^m = 80$ 。ここで、

$\mathbb{I}$  は全ての要素が 1 のベクトルを表す。始めの三つのハイパーパラメータ  $N_Z$ ,  $N_D$ ,  $N_K$  は図 4 のようなアノテーションデータの統計情報を参考にして決定した。データによると、多くの楽曲ではセクションの種類数は 12 以下であり、その継続長は 40 ビート以下である。あるセクションの継続長が 32 ビート（8 小節）であり、各コードの長さが 2 ビートであるとする、そのセクションにおけるコードの数として 16 が得られる。また、簡単のため  $\sigma$  の値は 1,  $w_{dur}$  の値は 4 とした。  $w_{dur}$  の値による性能への影響については後述する。その他のハイパーパラメータは後述の評価尺度に関して大まかに最適化を行うことで決定した。各パラメータはそのほかのパラメータを固定した上で、グリッドサーチによって最適化した。ハイパーパラメータの更なる最適化は今後の課題とした。

本研究では、三種類の実験を行なった。最初の実験では、三つの基本的側面をモデルに組み込んだことによる性能向上を検証した。音色の同質性とコード進行の反復性による影響を分析するため、提案法他、クロマのみを用いた場合と MFCC のみを用いた場合を比較した。さらに、継続長の規則性による影響を確認するため、継続長分布を一様分布とした場合も比較した。次の実験では、ビタビ学習による性能向上を確認するため、提案法とビタビ学習を行わない場合を比較した。最後の実験では、既存の手法との比較実験として、MSAF（音楽構造解析フレームワーク）[33] に実装されている、VMO（可変マルコフオラクル）[34]、CNMF（凸型 NMF）[15]、および SCluster（スペクトラルクラスタリング法）[16] などの最新の手法を用いた。これらのモデルでは、MSAF のデフォルトの設定を用いた。

評価実験では、セグメンテーションとクラスタリングの性能を MIREX [35] と同様の評価法により測定した。セグメンテーションの性能は、セクション境界位置に対する F 値  $F_{0.5}$ ,  $F_{0.5}^{(0.58)}$ ,  $F_{3.0}$ ,  $F_{3.0}^{(0.58)}$  により評価した。推定された境界は、 $F_{0.5}$ ,  $F_{0.5}^{(0.58)}$  では  $\pm 0.5$  秒,  $F_{3.0}$ ,  $F_{3.0}^{(0.58)}$  では  $\pm 3.0$  秒の区間に正解データの境界がある場合に正しいと判断した。適合率は正しく推定された境界の割合、再現率は正解の境界のうち正しく推定されたものの割合である。F 値  $F_{0.5}$ ,  $F_{3.0}$  は適合率と再現率の調和平均として定義する。また、 $F_{0.5}^{(0.58)}$ ,  $F_{3.0}^{(0.58)}$  は人間の音楽構造の認知と近いとされる評価尺度 [36] である。

$$F_{0.5}^{(0.58)} = (1 + 0.58^2) \frac{P_{0.5} R_{0.5}}{0.58^2 P_{0.5} + R_{0.5}} \quad (34)$$

$$F_{3.0}^{(0.58)} = (1 + 0.58^2) \frac{P_{3.0} R_{3.0}}{0.58^2 P_{3.0} + R_{3.0}} \quad (35)$$

ここで、 $P_{0.5}, R_{0.5}, P_{3.0}, R_{3.0}$  はそれぞれ正解とする幅を  $\pm 0.5$  秒とした時の適合率と再現率、および正解とする幅を  $\pm 3.0$  秒とした時の適合率と再現率である。

クラスタリングの性能は、次のように定義される F 値  $F_{\text{pair}}$  [37] により評価した。セクション構造の推定結果の

表 1 提案法の性能に関する比較実験の評価結果. 最下段は提案法である.

特徴量 クロマ MFCC	継続長分布	セグメンテーション				クラスタリング
		$F_{0.5}$ (%)	$F_{0.5}^{(0.58)}$ (%)	$F_{3.0}$ (%)	$F_{3.0}^{(0.58)}$ (%)	$F_{\text{pair}}$ (%)
✓	一様分布	7.50	8.16	31.3	34.3	27.9
	✓ 一様分布	24.1	25.0	60.8	62.9	56.5
✓	✓ 一様分布	21.9	22.2	56.3	57.2	52.3
✓	経験分布	17.1	17.9	38.9	40.4	29.7
	✓ 経験分布	<b>39.7</b>	<b>39.9</b>	<b>67.8</b>	<b>68.1</b>	<b>58.0</b>
✓	✓ 経験分布	33.0	32.9	58.7	58.2	54.3

表 2 ビタビ学習に関する比較実験の評価結果

ビタビ学習	セグメンテーション				クラスタリング
	$F_{0.5}$ (%)	$F_{0.5}^{(0.58)}$ (%)	$F_{3.0}$ (%)	$F_{3.0}^{(0.58)}$ (%)	$F_{\text{pair}}$ (%)
行わない	32.4	32.3	57.9	57.5	53.5
行う	<b>33.0</b>	<b>32.9</b>	<b>58.7</b>	<b>58.2</b>	<b>54.3</b>

中で同じセクションクラスに割り当てられたフレームのペアと、正解データの中で同様に同じセクションクラスに割り当てられたフレームのペアを比較した（フレーム長は 100 ms とする）。適合率  $P_{\text{pair}}$ 、再現率  $R_{\text{pair}}$ 、および F 値  $F_{\text{pair}}$  は以下のように定義する。

$$P_{\text{pair}} = \frac{|P_E \cap P_A|}{|P_E|}, \quad R_{\text{pair}} = \frac{|P_E \cap P_A|}{|P_A|} \quad (36)$$

$$F_{\text{pair}} = \frac{2P_{\text{pair}}R_{\text{pair}}}{P_{\text{pair}} + R_{\text{pair}}} \quad (37)$$

ここで、 $P_E$  は推定結果で同じクラスに割り当てられたフレームペアの集合、 $P_A$  は正解データで同じクラスに割り当てられたフレームペアの集合を示す。 $F_{0.5}^{(0.58)}$  と  $F_{3.0}^{(0.58)}$  を除くこれらの評価値は `mir_eval` ライブラリ [38] を用いて計算した。 $F_{0.5}^{(0.58)}$  と  $F_{3.0}^{(0.58)}$  は、`mir_eval` ライブラリを用いて得られた適合率と再現率を用いて計算した。

#### 4.2 実験結果

表 1 に提案法の性能に関する比較実験の結果を示す。継続長の規則性に関して、全ての場合において経験分布を用いることで性能が向上した。また、音色の同質性に関して、MFCC を用いる場合はどの評価値においても高い値を示した。一方、コード進行の反復性に関して、クロマのみを用いた場合は総じて MFCC のみを用いた場合よりも低い値を示しており、さらに MFCC と組み合わせた場合では MFCC のみを用いた場合よりも性能が低下した。この原因として、モデルが二層の潜在状態系列を持つことから、悪い局所解に陥りやすかった可能性が考えられる。各コードは 2 拍や 4 拍などの継続長を持つことが多いため、今後の課題として、下位マルコフ連鎖をセミマルコフモデルとすることでモデルを精密化すること、そしてパラメータ推定時に状態系列のサンプリングの数を増やして不良な局所解に陥るのを防ぐことが考えられる。これらの手法の改良を行う上で、計算量削減についても検討を進める必要

がある。

表 2 はビタビ学習に関する比較実験の結果を示している。ビタビ学習を行うことにより全体的に精度が向上した。このことから、ビタビ学習の有効性が確認された。

表 3 は既存手法との比較実験の結果を示している。SCluster は全ての評価尺度で、三つの既存手法の中では最も精度が高かった。VMO の推定結果の F 値は低く、推定結果には不自然に短いセグメントが多く見られた（図 4）。この結果は MSAF ライブラリの不適切な設定により生じたものであると考えられる。これら三つの手法と比較して、提案手法は全ての評価尺度で有意に優れた性能を示した。

#### 4.3 分析

最初に、各手法における推定結果の特徴をより詳しく調べた（図 4）。提案法の結果に対するセクション長の分布は、正解データに対する分布と特徴が類似している。特に、どちらの分布も 32 ビート（8 小節）と 16 ビートにおいてピークを持っている。一方で、比較手法に対する分布は正解データに対する分布と大きく異なっている。この結果は、セクションの継続長に関する規則性を捉えるモデリングの効果を明確に示している。同様に、セクション境界の拍節位置の分布においても、提案法の推定結果に対する分布と正解データに対する分布は類似している一方、比較手法の推定結果に対する分布はこれらと大きく異なっている。

正解データにおけるセクションクラス数は、おおよそ 8 から 12 の範囲に分布している。提案法における分布はやや低い方に移動しているものの、同様の形状をしている。この結果は、適切なセクションクラス数を自動で見出せるという提案法の非自明な能力を示すものであるが、実際よりもかなり小さいクラス数を推定することもしばしばあった。一方で、既存手法におけるセクションクラス数の分布ははるかにスパースであった。これらの手法の推定結果

表 3 既存手法との比較実験の評価結果

手法	セグメンテーション				クラスタリング
	$F_{0.5}$ (%)	$F_{0.5}^{(0.58)}$ (%)	$F_{3.0}$ (%)	$F_{3.0}^{(0.58)}$ (%)	$F_{\text{pair}}$ (%)
VMO [34]	7.71	5.32	8.52	5.88	28.5
CNMF [15]	6.61	6.82	37.4	38.6	41.7
SCluster [16]	13.6	14.4	50.4	51.8	45.5
提案法	<b>33.0</b>	<b>32.9</b>	<b>58.7</b>	<b>58.2</b>	<b>54.3</b>

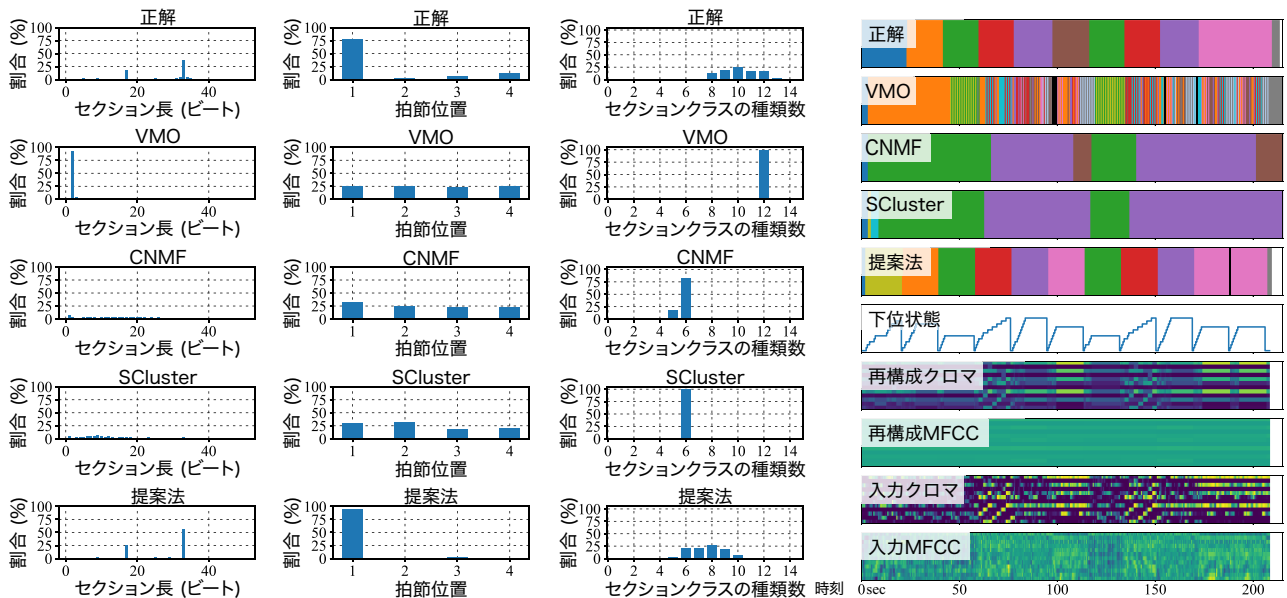


図 4 左側は、推定結果と正解データにおける、セクション長、セクション境界の拍節位置、およびセクションクラス数の分布を示す。右側は提案法および三つの比較手法による推定結果の例を示す (RWC-MDB-P-2001 No. 29)。下位状態系列はビタピアルゴリズムにより得られたものであり、再構成特徴量は対応する出力確率の期待値を示している。

は、全ての曲に対してほとんど同じセクションクラス数を持っていた。特に、CNMFとSClusterの推定結果は、正解データに比べクラス数が小さかった。

これらの解析により、提案法による音楽構造解析結果は、比較手法に比べて、人手による解析結果と近い特徴を持つことが確かめられた。また、これらの結果はF値をみるだけでは明らかにできないことは重要な点として指摘される。このことから、音楽構造解析の評価にはF値だけでは十分でないことがわかる。

図4に示す結果の例では、これらの傾向を観察できる。特に、CNMFとSClusterの推定結果は、セクションクラス数が小さく、セクション長も規則的でない。提案法の結果では、同じクラスのセクションでは下位モデルの潜在状態系列が似ていることが確認できる。これは提案法では、同じクラスのセクションにおいて、コード進行の反復を捉えられていることを示唆している。また、提案法の推定結果ではしばしば一部の低位状態のみを用いていることが見て取れる。提案法では低位のマルコフ連鎖の終了状態について制約を与えていないためどの状態でも終了することができ、極端な場合低位の初期状態のままそのセクションを

終えることが可能となっている。ここで、出力確率が無視できるほど自己遷移確率が他の遷移確率よりも非常に大きくなる場合、自己遷移を繰り返してしまい、その状態のままセクションが終了してしまうということが考えられる。これは低位のマルコフ連鎖に対する制約を増やすことで改善できる可能性がある。

さらに、提案法による音楽構造解析結果に含まれる典型的な誤りについて解析を行った。図5(a)では、同じクラスのセクションで、互いに継続長が大きく異なるものが見られる。一般に、ポピュラー音楽では同じクラスのセクションはほぼ同じ継続長を持つことが多く、この傾向を考慮したモデルを構築することでより正確な解析結果が得られると考えられる。また、図5(b)では、推定結果のセクション境界が全体的に少し前にずれている。このような誤りを軽減する方法として、新規性を扱い、セクション境界における特徴量の変化を捉えることが考えられる。

また、 $w_{\text{dur}}$ を{1, 4, 16, 64}の値で動かしたときの性能への影響を調査した。表4はその結果を示しており、この値の範囲で評価値が最大10%程度変化した。 $w_{\text{dur}}$ のより精密な最適化による多少の精度向上の可能性はあるが、その



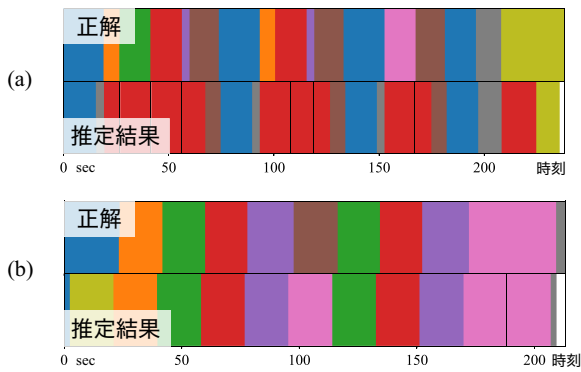


図 5 推定結果に含まれる典型的な誤り. (a) RWC-MDB-P-2001 No. 24. (b) RWC-MDB-P-2001 No. 29.

表 4  $w_{dur}$  による性能への影響

$w_{dur}$	1	4	16	64
セグメンテーション $F_{0.5}$ (%)	27.9	33.0	25.7	23.7

ほかのハイパーパラメータと同様に今後の課題とする。

#### 4.4 議論

本手法ではビートトラッキングの結果を入力の一部として用いているため、そこでの誤推定は本手法の推定結果に影響を与える。ビートトラッキングにおける典型的な誤りには推定テンポが正解の半分や倍となるものがあり、その場合、本手法により正解の半分や倍の長さを持つセクションが推定される可能性がある。しかし、最新のビートトラッキング手法の精度はポピュラー楽曲に対して 80% 後半から 90% 前半程度であり [39], ビートトラッキングでの誤りによる本手法の性能に対する影響は小さいと考えられる。また、セクション長の経験分布は図 4 中の正解データにおける分布に見るように、ピーク位置のビート数の半分や倍のビート数においてもある程度の確率値を持つ。そのため、ビートトラッキングで誤推定があっても、本手法ではセクションを正しく推定することがある程度は可能である。今後の拡張として、ビートトラッキングを前処理として使用せずに、音楽構造と同時にビート位置も推定することが考えられる。このような同時推定によって、典型的なセクション長のビート数の事前情報をビートトラッキングに取り入れることで、ビートトラッキングと構造解析の両方の精度が改善される可能性がある。

#### 5. 結論

本稿では、セクション内およびセクション間の構造を統一的に記述するベイズ HHSMM に基づく統計的音楽構造解析手法について論じた。音楽のセクションに関する三つの重要な側面である同質性、反復性、および規則性を取り入れたモデルを構築した。音楽のセグメンテーションとクラスタリングは教師なしベイズ学習に基づいて同時に行う

ことができ、音楽的に重要な特性である反復性やセクション長の分布をベイズ拡張により取り入れた。実験結果から、提案法により従来の代表的手法と比べて有意に優れたセグメンテーションとクラスタリングの精度が得られることを確認したが、反復性による性能向上は認められなかった。また、提案法による解析結果が人手による解析結果と類似する統計的性質を持つことを確かめた。

今後は反復性のモデルへの取り入れ方の改善を行うとともに、新規性の側面をモデルに取り込む改良を計画している。また、音楽は動機、フレーズ、セクション、セクショングループというように階層構造を持つ [40] ことから、より多くの階層を扱えるように拡張することも重要である [16]。教師なし学習に基づく提案法は、深層識別モデル [6-8] に基づく統計手法と相補的な関係にある。そこで、VAE (変分自己符号化) [22] の枠組みを用いてこれらのモデルを統合する方向性が有望であると考えられる。

謝辞 本研究の一部は、科研費 No. 19H04137, No. 19K20340, No. 16H01744 および JST ACCEL No. JPMJAC1602 の支援を受けた。

#### 参考文献

- [1] Paulus, J., Müller, M. and Klapuri, A.: State of the Art Report: Audio-Based Music Structure Analysis, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 625-636 (2010).
- [2] Foote, J.: Automatic Audio Segmentation Using a Measure of Audio Novelty, *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 452-455 (2000).
- [3] Jensen, K.: Multiple Scale Music Segmentation Using Rhythm, Timbre, and Harmony, *EURASIP Journal on Applied Signal Processing*, Vol. 2007, No. 1, pp. 159-159 (2007).
- [4] Serrà, J., Müller, M., Grosche, P. and Arcos, J.: Unsupervised Detection of Music Boundaries by Time Series Structure Features, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 1613-1619 (2012).
- [5] Peeters, G. and Bisot, V.: Improving Music Structure Segmentation Using Lag-Priors, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 337-342 (2014).
- [6] Ullrich, K., Schlüter, J. and Grill, T.: Boundary Detection in Music Structure Analysis Using Convolutional Neural Networks, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 417-422 (2014).
- [7] Grill, T. and Schlüter, J.: Music Boundary Detection Using Neural Networks on Combined Features and Two-Level Annotations, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 531-537 (2015).
- [8] Maezawa, A.: Music Boundary Detection Based on a Hybrid Deep Model of Novelty, Homogeneity, Repetition and Duration, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 206-210 (2019).

- [9] Sargent, G., Bimbot, F. and Vincent, E.: Estimating the Structural Segmentation of Popular Music Pieces Under Regularity Constraints, *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 25, No. 2, pp. 344–358 (2017).
- [10] Kaiser, F. and Peeters, G.: A Simple Fusion Method of State And Sequence Segmentation for Music Structure Discovery, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 257–262 (2013).
- [11] Aucouturier, J.-J. and Sandler, M.: Segmentation of Musical Signals Using Hidden Markov Models, *Audio Engineering Society (AES) Convention*, pp. 1–8 (2001).
- [12] Cooper, M. and Foote, J.: Summarizing Popular Music via Structural Similarity Analysis, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 127–130 (2003).
- [13] Goodwin, M. M. and Laroche, J.: A Dynamic Programming Approach to Audio Segmentation and Speech/Music Discrimination, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 309–312 (2004).
- [14] Grohganz, H., Clausen, M., Jiang, N. and Müller, M.: Converting Path Structures Into Block Structures Using Eigenvalue Decompositions of Self-Similarity Matrices, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 209–214 (2013).
- [15] Nieto, O. and Jehan, T.: Convex Non-negative Matrix Factorization for Automatic Music Structure Identification, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 236–240 (2013).
- [16] McFee, B. and Ellis, D. P. W.: Analyzing Song Structure with Spectral Clustering, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 405–410 (2014).
- [17] Ren, L., Dunson, D., Lindroth, S. and Carin, L.: Dynamic Nonparametric Bayesian Models for Analysis of Music, *Journal of the American Statistical Association (JASA)*, Vol. 105, No. 490, pp. 458–472 (2008).
- [18] Barrington, L., Chan, A. B. and Lanckriet, G.: Modeling Music as a Dynamic Texture, *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 18, No. 3, pp. 602–612 (2010).
- [19] Goto, M.: A Chorus Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station, *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, Vol. 14, No. 5, pp. 1783–1794 (2006).
- [20] Paulus, J. and Klapuri, A.: Music Structure Analysis Using a Probabilistic Fitness Measure and a Greedy Search Algorithm, *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 17, No. 6, pp. 1159–1170 (2009).
- [21] Cheng, T., Smith, J. B. L. and Goto, M.: Music Structure Boundary Detection and Labelling by a Deconvolution of Path-Enhanced Self-Similarity Matrix, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 106–110 (2018).
- [22] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes, *International Conference on Learning Representations (ICLR)*, pp. 1–14 (2014).
- [23] Shibata, G., Nishikimi, R., Nakamura, E. and Yoshii, K.: Statistical Music Structure Analysis Based on a Homogeneity-, Repetitiveness-, and Regularity-Aware Hierarchical Hidden Semi-Markov Model, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 268–275 (2019).
- [24] Smith, J. B. L. and Goto, M.: Using Priors to Improve Estimates of Music Structure, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 554–560 (2016).
- [25] Levy, M. and Sandler, M.: New methods in structural segmentation of musical audio, *European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (2006).
- [26] Allahverdyan, A. and Galstyan, A.: Comparative Analysis of Viterbi training and Maximum Likelihood Estimation for HMMs, *Advances in Neural Information Processing Systems (NIPS)*, pp. 1674–1682 (2011).
- [27] Yu, S.-Z.: Hidden Semi-Markov Models, *Artificial Intelligence*, Vol. 174, No. 2, pp. 215–243 (2010).
- [28] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases, *International Conference on Music Information Retrieval (ISMIR)*, pp. 287–288 (2002).
- [29] Goto, M.: AIST Annotation for the RWC Music Database, *International Conference on Music Information Retrieval (ISMIR)*, pp. 359–360 (2006).
- [30] Wu, Y. and Li, W.: Automatic Audio Chord Recognition With MIDI-Trained Deep Feature and BLSTM-CRF Sequence Decoding Model, *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 27, No. 2, pp. 355–366 (2019).
- [31] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E. and Nieto, O.: librosa: Audio and Music Signal Analysis in Python, *Python in Science Conference*, pp. 18–24 (2015).
- [32] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F. and Widmer, G.: madmom: A New Python Audio and Music Signal Processing Library, *ACM International Conference on Multimedia (ACMMM)*, pp. 1174–1178 (2016).
- [33] Nieto, O. and Bello, J. P.: Systematic Exploration Of Computational Music Structure Research, *International Society for Music Information Retrieval Conference (ISMIR)* (2016).
- [34] Wang, C.-i. and Mysore, G. J.: Structural Segmentation with the Variable Markov Oracle and Boundary Adjustment, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 291–295 (2016).
- [35] Ehmann, A. F., Bay, M., Downie, J. S., Fujinaga, I. and Roure, D. D.: Music Structure Segmentation Algorithm Evaluation: Expanding on MIREX 2010 Analyses and Datasets, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 561–566 (2011).
- [36] Nieto, O., Farbood, M. M., Jehan, T. and Bello, J. P.: Perceptual analysis of the f-measure for evaluating section boundaries in music, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 265–270 (2014).
- [37] Levy, M. and Sandler, M.: Structural Segmentation of Musical Audio by Constrained Clustering, *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, Vol. 16, No. 2, pp. 318–326 (2008).
- [38] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D. and Ellis, D. P. W.: mir\_eval: A

Transparent Implementation of Common MIR Metrics, *International Society for Music Information Retrieval Conference (ISMIR)* (2014).

- [39] Böck, S., Krebs, F. and Widmer, G.: Joint Beat and Downbeat Tracking with Recurrent Neural Networks, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 255-261 (2016).
- [40] Lerdahl, F. and Jackendoff, R.: *A Generative Theory of Tonal Music*, MIT Press (1983).




**柴田 剛** (学生会員)

2019年京都大学工学部情報学科卒業。同年より同大学大学院情報学研究科知能情報学専攻修士課程在学。音楽情報処理の研究に従事。




**錦見 亮** (学生会員)

2018年京都大学大学院情報学研究科知能情報学専攻修士課程修了。同年より同専攻博士後期課程在学。音楽情報処理の研究に従事。2017年第116回音楽情報科学研究会 学生奨励賞, 平成30年度 山下記念研究賞受賞。



**中村 栄太** (正会員)

2012年東京大学大学院理学系研究科物理学専攻博士課程修了。博士(理学)。国立情報学研究所, 明治大学, 京都大学などで研究員を経て, 2019年から京都大学白眉センター特定助教。音楽知能情報の研究に従事。



**吉井 和佳** (正会員)

2008年京都大学大学院情報学研究科博士後期課程修了。同年産業技術総研究所情報技術研究部門に入所。2018年京都大学大学院情報学研究科准教授に就任。音楽情報処理, 統計的音響信号処理の研究に従事。博士(情報学)。