

LSTM-HSMM ハイブリッドモデルに基づく音楽構造解析

柴田 剛^{1,a)} 錦見 亮^{1,b)} 中村 栄太^{1,c)} 吉井 和佳^{1,d)}

概要：音楽構造解析では、セクションの境界推定と各セクションに対するイントロ・A メロ・B メロ・サビなどのラベル付けを行う。本稿では、LSTM-HSMM ハイブリッドモデルに基づくポピュラー音楽に対する構造解析手法を提案する。本モデルはセグメンテーション・クラスタリングを行う階層 HSMM（隠れセミマルコフモデル）とラベリングを行う LSTM ネットワークの統合モデルであり、クロマ特徴量・MFCC・メルスペクトルの生成過程を表す。セクション系列とクロマ特徴量・MFCC の生成確率は入力 of 音響特徴量から学習することで、音楽セクションの持つ重要な性質である同質性・反復性・規則性を明示的に表現する。LSTM ネットワークはセクションラベルに対するメルスペクトルの生成確率を記述し、事前学習を行うことで各セクションラベルの一般的な特徴を捉える。最終的にピタビ復号により最尤のセクションラベル系列を推定する。実験により提案手法の有効性を検証する。

1. はじめに

イントロ・A メロ・B メロ・サビなどの複数のセクションからなる音楽構造はポピュラー音楽における主要な要素の一つであり、音楽構造解析は音楽情報検索（Music Information Retrieval; MIR）[1] の基礎技術である。多くの研究では、音響信号をいくつかのセクションに分割するセグメンテーションタスク [2–9]、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タスク [10–18]、またはその両方に取り組んできた。クラスタリングタスクでは「A」や「B」などのような任意のラベルを付与するが、「A メロ」や「B メロ」、「サビ」のような慣習的に用いられるラベルは再生ナビゲーションに有用であるため [19]、本研究ではこれらのラベルを付与するラベリングタスクにも取り組む [20, 21]。

音楽構造解析では、音響特徴量の同質性と反復性、セクション長の規則性、およびセクション境界での新規性がセクションにおける 4 つの主要な性質とされてきた。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、クロマ特徴量やコード進行）が同じラベルのセクションで繰り返されることを意味する。規則性は、例え

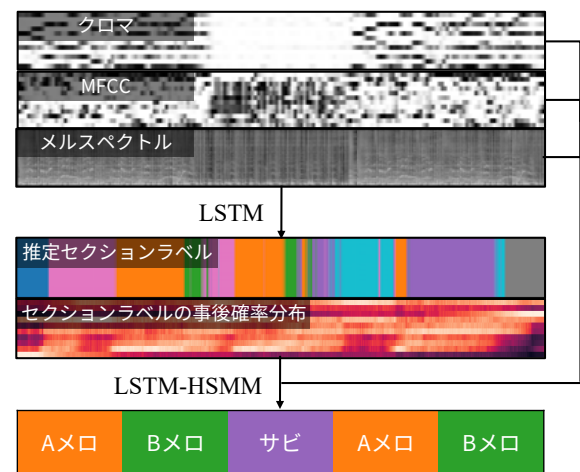


図 1 提案する音楽構造解析手法

ばポピュラー音楽では、セクション長の多くが 4 または 8 小節のように規則的であることを意味する。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。我々は、これまでにセグメンテーションとクラスタリングを統一的に行う階層隠れセミマルコフモデル（Hidden Semi-Markov Model; HSMM）[22] を提案してきた。本モデルは先に述べた 4 つの性質のうち、同質性・反復性・規則性を明示的に表現し、観測特徴量から潜在セクション系列を推論できる確率的生成モデルである。しかし、この手法はセクションのクラスターは推定できるが、音楽構造解析を完了するにはそれぞれのクラスターに具体的なラベルを与えなければならない。

ラベリングでは音響特徴量とセクションラベルの複雑な関係性を捉える必要がある。これには、複雑な分類問題に

¹ 京都大学大学院情報学研究科

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

広く用いられる深層ニューラルネットワーク (Deep Neural Network; DNN) の適用が考えられる。単純には、先述した階層 HSMM で得られたクラスターごとに DNN で得られたラベルを多数決法などを用いて割り当てることでラベリングを達成できる。しかし、この方法ではセグメンテーションとクラスタリングでの誤りが伝播し、ラベリングにおいてその誤りを修正することができない。DNN によって得られた情報はセグメンテーションとクラスタリングでも有用であると考えられるため、DNN によるラベル推定と階層 HSMM によるセクション構造推定を統合する方法が有効であろう。

本稿では、双方向長短期記憶 (Bidirectional Long Short-Term Memory; BLSTM) ネットワークのラベリング能力を階層 HSMM を用いた生成フレームワークに統合する音楽構造解析手法を提案する (図 1)。本モデルは、各セクション内で均質である MFCC, 同一ラベルのセクション同士で繰り返されるクロマ特徴量, および規則的な継続時間長を持つセクションから得られるメルスペクトルの生成過程を表現する。BLSTM ネットワークをフレーム単位のメルスペクトルからセクションラベルを推定するように教師あり学習し、ネットワークから得られたセクションラベルの事後確率分布とセクションラベルの経験的事前分布からメルスペクトルの出力確率分布を実行時に計算する。ギブスサンプリングおよびそれに続くビタビ復号を用いることで、音響特徴量が与えられると、セクションの初期確率、遷移確率、継続長確率、終了確率と同様に潜在セクション系列を推定する。悪い局所解を避けるため、潜在セクション系列はネットワークからの出力を用いて初期化する。

本研究の主な貢献は、各楽曲に適応可能なベイズ LSTM-HSMM ハイブリッドモデルに基づき、セグメンテーション・ラベリングを統一的に行う統計的手法の提案である。セクションの統計的特徴は楽曲ごとに固有であるため、それらの経験的事前分布に基づくベイズ推定は音楽構造解析の性能向上において重要な役割を果たしている。我々は、後処理的に BLSTM のラベリング結果を用いたカスケードモデル, およびビタビ復号のみを用いたベースラインモデルと比較して、提案手法が優れていることを実験的に示した。

2. 関連研究

本章では、音楽構造解析の手法をセグメンテーション、クラスタリング, およびラベリングの観点から総括する。

2.1 セグメンテーション

セグメンテーションでは、新規性が中心的な役割を果たす。Foote [2] はチェッカーボードカーネルを自己類似度行列 (Self-Similarity Matrix; SSM) の対角要素に沿って畳み込むことで得られた時変新規性曲線を用いてピーク

検出を行った。Jensen [3] は同質性と新規性に基づくコスト関数を最小化することでセクション境界の検出を行った。Goto [20] および Serrà [4] は反復構造が垂直線として現れるラグ SSM から計算される新規性曲線を提案した。これらの手法は Peeters ら [5] によって統合され、また Serrà らの手法 [4] はクラスタリングを取り扱うよう拡張された [23]。最近では、Ullrich ら [6] が畳み込みニューラルネットワークに基づく教師あり手法を提案しており、これは粗い境界アノテーションと細かい境界アノテーションを扱う手法に拡張されている [7]。Smith ら [24] は後処理的ではなく、解析時に規則性を考慮することの重要性を強調した。Sargent ら [9] はセクション長の規則性の有効性を指摘した。Maezawa [8] は同質性・反復性・新規性および規則性を考慮するコスト関数に基づいた、LSTM ネットワークを用いる手法を提案した。

2.2 クラスタリング・ラベリング

Cooper ら [12] はセグメンテーション [2] とセクション内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 から非対角成分中の繰り返しの縞構造を効率的に検出した。反復性と同質性を取り扱うため、Grohganz ら [14] は非対角成分の縞構造を固有値分解を用いてブロック対角構造に変換した。Nieto ら [15] は凸結合の制約を加えた非負値行列因子分解を用いてセグメンテーションとクラスタリングを行った。McFee ら [16] は反復構造をグラフに変換し、グラフ分解のためのスペクトルクラスタリングを行った。Cheng ら [25] は Grohganz ら [14] と同様に、NMFD (Nonnegative Matrix Factor Deconvolution) を用いてパス強調 SSM をブロック強調 SSM に変換した。

セグメンテーションとクラスタリングを統一的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 を用いた手法を行った。Levy ら [26] はセクション長の規則性に基づく HSMM を提案した。Ren ら [17] は適切なセクション数を推定できる HMM のノンパラメトリックベイズ拡張を提案した。Barrington ら [18] は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た。

セクションラベルを推定する試みをした研究は少ない。Maddage ら [27] は典型的な音楽構造と各セクションの役割に基づくラベリング手法を提案した。Paulus ら [21] はセクションの N グラムに基づく確率的適合度関数を用いてセグメンテーションとラベリングを行った。

3. 提案手法

本章では提案する音楽構造解析手法について説明する。

3.1 問題設定

本稿で我々が取り組む問題を以下のように定式化する。 T は四分音符単位のビート数, N は曲中のセクション数, K はセクションラベル数, L はセクション中のビート数の最大値とする。

仮定: 音楽音響信号のビート時刻はあらかじめビートトラッキング手法 [28] で推定されているものとする。

入力: 音楽音響信号から得られた, ビート単位のクロマ特徴量系列 $\mathbf{X}^c \triangleq \mathbf{x}_{1:T}^c$ ($\mathbf{x}_t^c \in \mathbb{R}^{12}$), MFCC 系列 $\mathbf{X}^m \triangleq \mathbf{x}_{1:T}^m$ ($\mathbf{x}_t^m \in \mathbb{R}^{12}$), およびメルスペクトル系列 $\mathbf{X}^s \triangleq \mathbf{x}_{1:T}^s$ ($\mathbf{x}_t^s \in \mathbb{R}^{128}$)

出力: セクションラベル系列 $\mathbf{Z} \triangleq z_{1:N}$ ($z_n \in \{1, \dots, K\}$) とそれらの継続長 $\mathbf{D} \triangleq d_{1:N}$ ($d_n \in \{1, \dots, L\}$)

ここで, MFCC は低次の 12 次元を用い, メルスペクトルの周波数ビン数は 128 を用いる。また, $i:j$ は i から j までの添字集合を表す。以後, $\{\mathbf{X}^c, \mathbf{X}^m, \mathbf{X}^s\}$ を $\mathbf{X}$, $\{\mathbf{x}^c, \mathbf{x}^m, \mathbf{x}^s\}$ を $\mathbf{x}$ と表記する。

3.2 モデル定式化

図 2 に示すように, セクションラベルとコードラベルの二つの潜在系列からなる階層 HSMm を定式化する。 $\mathbf{S}_n \triangleq s_{n,1:d_n}$ ($s_{n,\tau} \in \{1, \dots, M\}$) をセクション n におけるコード系列とし, $\mathbf{S} \triangleq \mathbf{S}_{1:N}$ をコード系列 $\mathbf{S}_n$ の系列, M をセクション中のコード数の最大値とし, 確率 $p(\mathbf{X}, \mathbf{Z}, \mathbf{D}, \mathbf{S})$ を以下のように定義する。

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{D}, \mathbf{S}) = p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S})p(\mathbf{S}|\mathbf{Z}, \mathbf{D})p(\mathbf{Z}, \mathbf{D}) \quad (1)$$

ここで, $p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S})$ は観測特徴量 $\mathbf{X}$ に関する音響モデル, $p(\mathbf{S}|\mathbf{Z}, \mathbf{D})$ はコード系列 $\mathbf{S}$ に関する Left-to-Right 型マルコフモデル, $p(\mathbf{Z}, \mathbf{D})$ はセクションラベル系列 $\mathbf{Z}$ およびその継続長 $\mathbf{D}$ に関する全遷移型セミマルコフモデルである。

3.2.1 セクションラベルに関するセミマルコフ連鎖

式 (1) 中の全遷移型セミマルコフモデル $p(\mathbf{Z}, \mathbf{D})$ は, 以下のようにセクションラベル系列 $\mathbf{Z}$ およびその継続長 $\mathbf{D}$ の生成過程を表現する。

$$p(\mathbf{Z}, \mathbf{D}) = p(z_1, d_1) \prod_{n=2}^N p(z_n, d_n | z_{n-1}, d_{n-1}) \quad (2)$$

それぞれの項は以下で与える。

$$p(z_1, d_1) = \rho_{z_1} \psi_{d_1} \quad (3)$$

$$p(z_n, d_n | z_{n-1}, d_{n-1}) = \pi_{z_{n-1}z_n} \psi_{d_n} \quad (4)$$

$$p(z_N, d_N | z_{N-1}, d_{N-1}) = \pi_{z_{N-1}z_N} \psi_{d_N} v_{z_N} \quad (5)$$

ここで, ρ_z , $\pi_{zz'}$, および v_z はそれぞれセクションラベルの初期確率, 遷移確率, 終了確率であり, ψ_d は継続長確率である。

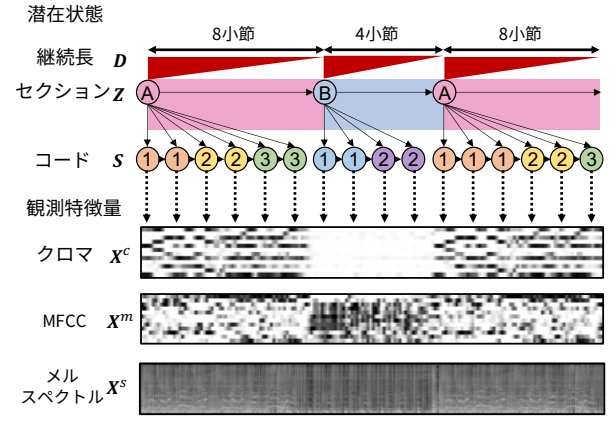


図 2 提案する LSTM-HSMm ハイブリッドモデル

3.2.2 コードに関する Left-to-Right 型マルコフモデル

式 (1) 中の Left-to-Right 型マルコフモデル $p(\mathbf{S}|\mathbf{Z}, \mathbf{D})$ は, 以下のようにコードラベル系列 $\mathbf{S}$ の生成過程を表現する。

$$p(\mathbf{S}|\mathbf{Z}, \mathbf{D}) = \prod_{n=1}^N p(s_{n,1}) \prod_{\tau=2}^{d_n} p(s_{n,\tau} | s_{n,\tau-1}, z_n) \quad (6)$$

それぞれの項は以下で与える。

$$p(s_{n,1} = 1) = 1 \quad (7)$$

$$p(s_{n,\tau} | s_{n,\tau-1}, z_n) = \phi_{s_{n,\tau-1}s_{n,\tau}}^{(z_n)} \quad (8)$$

ここで, z_n は対応するセクションのラベル, $\phi_{ss'}^{(z)}$ は状態 s から状態 s' への遷移確率である。Left-to-Right 型マルコフモデルは初期状態は $s_{n,1} = 1$ であり, $\tau_1 < \tau_2$ のとき $s_{n,\tau_1} \leq s_{n,\tau_2}$ を満たす。また, 遷移においてスキップ可能な状態数の最大値を表すハイパーパラメータ σ を導入する。すなわち, 状態 s から状態 $s + \sigma$ への遷移を許すことで, 同じラベルのセクション同士でコード進行にいくらかのバリエーションを持たせることができる。

3.2.3 音響特徴量の出力

クロマ特徴量 $\mathbf{X}^c$, MFCC $\mathbf{X}^m$, およびメルスペクトル $\mathbf{X}^s$ が条件付き独立かつ時間的に独立であるとする, 式 (1) 中の音響モデル $p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S})$ は以下のように分解できる。

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S}) = \prod_{t=1}^T \chi_{z_t, s_t}^c(\mathbf{x}_t^c) \chi_{z_t}^m(\mathbf{x}_t^m) \chi_{z_t}^s(\mathbf{x}_t^s) \quad (9)$$

ここで, z_t と s_t はそれぞれビート t におけるセクションラベルとコードであり, $\chi_{z,s}^c$, χ_z^m , χ_z^s はそれぞれクロマ特徴量 $\mathbf{x}^c$, MFCC $\mathbf{x}^m$, およびメルスペクトル $\mathbf{x}^s$ の出力確率である。

クロマ特徴量 $\mathbf{x}^c \in \mathbb{R}^{12}$ はセクションラベル z とコード s の両方に依存して生成する。このようにして, コードの反復性を同じラベルの全てのセクションに同一の出力確率分布を適用することで表現する。式 (9) 中の出力確率 $\chi_{z,s}^c(\mathbf{x}^c)$ は以下のように多変量正規分布で与える。

$$\chi_{z,s}^c(\mathbf{x}^c) = \mathcal{N}(\mathbf{x}^c | \boldsymbol{\mu}_{z,s}^c, (\boldsymbol{\Lambda}_{z,s}^c)^{-1}) \quad (10)$$

ここで、 $\boldsymbol{\mu}_{z,s}^c$ と $\boldsymbol{\Lambda}_{z,s}^c$ はそれぞれ平均ベクトルおよび精度行列である。

MFCC $\mathbf{x}^m \in \mathbb{R}^{12}$ はセクションラベル z に依存して生成する。これにより、モデルがそれぞれのセクションにおける音色特徴量の同質性を捉えることができる。式 (9) 中の出力確率 $\chi_z^m(\mathbf{x}^m)$ は以下のように多変量正規分布で与える。

$$\chi_z^m(\mathbf{x}^m) = \mathcal{N}(\mathbf{x}^m | \boldsymbol{\mu}_z^m, (\boldsymbol{\Lambda}_z^m)^{-1}) \quad (11)$$

ここで、 $\boldsymbol{\mu}_z^m$ と $\boldsymbol{\Lambda}_z^m$ はそれぞれ平均ベクトルおよび精度行列である。

メルスペクトル $\mathbf{x}^s \in \mathbb{R}^{128}$ はセクションラベル z に依存して生成する。式 (9) 中の出力確率 $\chi_z^s(\mathbf{x}^s)$ は以下のように計算する。

$$\chi_z^s(\mathbf{x}^s) = p(\mathbf{x}^s | z) \propto \frac{p(z | \mathbf{x}^s)}{p(z)} \quad (12)$$

ここで、 $p(z)$ はセクションラベルのユニグラム確率、 $p(z | \mathbf{x}^s)$ はフレーム単位のメルスペクトルからセクションラベルを推定する BLSTM から得る。 $p(z | \mathbf{x}^s)$ は BLSTM のフレーム単位の出力をビート単位に平均化したものとする。

3.2.4 音楽知識に基づく事前分布

音楽セクションに関する事前知識を利用するため、モデルパラメータ $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\rho}, \boldsymbol{\psi}, \boldsymbol{\pi}, \boldsymbol{v}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ に共役事前分布をおくことでベイズ HSMM を定式化する。多変量正規分布に対してガウス・ウィシャート事前分布をおく。

$$\begin{aligned} \boldsymbol{\mu}_{z,s}^c, \boldsymbol{\Lambda}_{z,s}^c &\sim \mathcal{N}(\boldsymbol{\mu}_{z,s}^c | \mathbf{m}_0^c, (\beta_0^c \boldsymbol{\Lambda}_{z,s}^c)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_{z,s}^c | \mathbf{W}_0^c, \nu_0^c) \\ \boldsymbol{\mu}_z^m, \boldsymbol{\Lambda}_z^m &\sim \mathcal{N}(\boldsymbol{\mu}_z^m | \mathbf{m}_0^m, (\beta_0^m \boldsymbol{\Lambda}_z^m)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_z^m | \mathbf{W}_0^m, \nu_0^m) \end{aligned}$$

ここで、 $\mathbf{m}_0^c$, β_0^c , $\mathbf{W}_0^c$, ν_0^c , $\mathbf{m}_0^m$, β_0^m , $\mathbf{W}_0^m$ および ν_0^m はハイパーパラメータである。次に、離散分布に対してディリクレ事前分布をおく。

$$\boldsymbol{\rho} \triangleq \rho_{1:K} \sim \text{Dirichlet}(\mathbf{a}^\rho) \quad (13)$$

$$\boldsymbol{\psi} \triangleq \psi_{1:L} \sim \text{Dirichlet}(\mathbf{a}^\psi) \quad (14)$$

$$\boldsymbol{\pi}_z \triangleq \pi_{z(1:K)} \sim \text{Dirichlet}(\mathbf{a}^{\pi_z}) \quad (15)$$

$$\boldsymbol{v} \triangleq v_{1:K} \sim \text{Dirichlet}(\mathbf{a}^v) \quad (16)$$

$$\boldsymbol{\phi}_s^{(z)} \triangleq \phi_{s(1:M)}^{(z)} \sim \text{Dirichlet}(\mathbf{a}^\phi) \quad (17)$$

ここで、 $\mathbf{a}^\rho$, $\mathbf{a}^\psi$, $\mathbf{a}^{\pi_z}$, $\mathbf{a}^v$ および $\mathbf{a}^\phi$ はハイパーパラメータである。ベイズ推論の重要な利点として、これらのスパース性に関わるハイパーパラメータを制御することで不要なセクションを自動的に除去できる。

ポピュラー音楽において、「B メロ」は「A メロ」の後に来やすく、セクション長は 4 小節の整数倍になりやすい。このような統計的傾向は事前分布に組み込むことがで

きる。具体的には、 $\mathbf{a}^\rho$ としてセクションの経験的初期確率 $\mathbf{a}_{\text{emp}}^\rho$, $\mathbf{a}^{\pi_z}$ としてセクションの経験的遷移確率 $\mathbf{a}_{\text{emp}}^{\pi_z}$, $\mathbf{a}^v$ としてセクションの経験的終了確率 $\mathbf{a}_{\text{emp}}^v$, $\mathbf{a}^\psi$ としてセクションの経験的継続長確率 $\mathbf{a}_{\text{emp}}^\psi$ をそれぞれ定数倍したものを用いる。

3.3 ベイズ推論

事後分布 $p(\mathbf{Z}, \mathbf{D}, \mathbf{S}, \boldsymbol{\Theta} | \mathbf{X})$ は解析的に扱いにくいいため、ギブスサンプリング法を用いる。最初に分布 $p(\mathbf{Z}, \mathbf{D}, \mathbf{S} | \boldsymbol{\Theta}, \mathbf{X})$ から潜在変数 $\mathbf{Z}$, $\mathbf{D}$, $\mathbf{S}$ をサンプルしたのち、分布 $p(\boldsymbol{\Theta} | \mathbf{Z}, \mathbf{D}, \mathbf{S}, \mathbf{X})$ からモデルパラメータ $\boldsymbol{\Theta}$ をサンプルする。

3.3.1 事前学習

学習データから経験分布 $\mathbf{a}_{\text{emp}}^\rho$, $\mathbf{a}_{\text{emp}}^\psi$, $\mathbf{a}_{\text{emp}}^{\pi_z}$, および $\mathbf{a}_{\text{emp}}^v$ を計算する。 $(\mathbf{a}_{\text{emp}}^\rho)_z$ はセクションラベル系列がセクション z から始まる回数、 $(\mathbf{a}_{\text{emp}}^\psi)_d$ はセクションが継続長 d を持つ回数、 $(\mathbf{a}_{\text{emp}}^{\pi_z})_{z'}$ はセクション z からセクション z' への遷移回数、 $(\mathbf{a}_{\text{emp}}^v)_z$ はセクションラベル系列がセクション z で終わる回数である。

3.3.2 潜在変数の初期化

悪い局所解を避けるため、BLSTM を用いてセクションラベル $\mathbf{Z}$ と継続長 $\mathbf{D}$ を初期化する。BLSTM によって推定されたセクションラベルのフレーム単位での事後確率は各ビート t で平均化する。各ビートのセクションラベルは、そのビートで最大事後確率を得るラベルとする。連続するセクションラベルは単一のセクションとみなすが、継続長が 4 拍 (1 小節) よりも短い場合、そのセクションは事後確率に応じて左か右のセクションに統合する。長すぎる区間を扱うのは効率が悪いいため、32 拍以上の継続長を持つセクションを 32 拍ごとに分割する。その後、コード系列 $\mathbf{S}$ とモデルパラメータ $\boldsymbol{\Theta}$ をサンプルする。

3.3.3 潜在変数のサンプリング

潜在変数 $\mathbf{Z}$, $\mathbf{D}$, $\mathbf{S}$ 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 $t - d_t + 1$ で始まり、ビート t で終わるセクションのラベルと継続長を表す変数 z_t , d_t を導入する。また、このセクションのコードに関する周辺出力確率を $\omega_{z_t}(\mathbf{x}_{t-d_t+1:t})$ とする。この確率は、コードに関するマルコフモデルに対してフォワードアルゴリズムを用いることで計算する。

セクションラベルに関するマルコフモデルのフォワードフィルタリングステップでは、変数 $\alpha_t(z_t, d_t) = p(z_t, d_t, \mathbf{x}_{1:t})$ の初期化と更新を行う。

$$\alpha_t(z_t, d_t = t) = \rho_{z_t} \psi_{d_t} \omega_{z_t}(\mathbf{x}_{1:t}) \quad (18)$$

$$\begin{aligned} \alpha_t(z_t, d_t) &= \sum_{z', d'} \alpha_{t-d_t}(z', d') \pi_{z' z_t} \psi_{d_t} \omega_{z_t}(\mathbf{x}_{t-d_t+1:t}) \quad (19) \end{aligned}$$

バックワードサンプリングステップでは、セクションラベル $\mathbf{Z}$ と継続長 $\mathbf{D}$ を逆順にサンプルする。

$$p(z_T, d_T | \mathbf{X}) \propto \alpha_T(z_T, d_T) \quad (20)$$

変数 z_t および d_t がサンプルされたとき、ビート $t' = t - d_t$ における変数 $z_{t'}$ と $d_{t'}$ は以下の確率にしたがってサンプルする。

$$p(z_{t'}, d_{t'} | z_{t:T}, d_{t:T}, \mathbf{X}) \propto \alpha_{t'}(z_{t'}, d_{t'}) \pi_{z_{t'}, z_t} \quad (21)$$

次に、コード系列 $\mathbf{S}$ をサンプル済みの $\mathbf{Z}$ と $\mathbf{D}$ を用いてサンプルする。それぞれのコード系列 $\mathbf{S}_n$ はセクション n におけるコードに関するマルコフモデルにおける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てサンプルする。フォワードフィルタリングステップでは、確率 $\zeta_{n, s_n, \tau}$ を再帰的に計算する。

$$\begin{aligned} \zeta_{n, s_n, 1} &= p(s_{n,1}, \mathbf{x}_1 | z_n, d_n) \\ &= \delta_{s_{n,1}, 1} \chi_{z_n, 1}(\mathbf{x}_1) \end{aligned} \quad (22)$$

$$\begin{aligned} \zeta_{n, s_n, \tau} &= p(s_{n,\tau}, \mathbf{x}_{1:\tau} | z_n, d_n) \\ &= \left(\sum_{s_{n,\tau-1}} \zeta_{n, s_{n,\tau-1}} \phi_{s_{n,\tau-1}, s_{n,\tau}}^{(z_n)} \right) \chi_{z_n, s_n, \tau}(\mathbf{x}_\tau) \end{aligned} \quad (23)$$

ここで、 $\mathbf{x}_\tau$ はセクション境界に対して相対的に定めるビート $\tau \in \{1, \dots, d_n\}$ における観測特徴量、 $\chi_{z,s}(\mathbf{x})$ は周辺確率 $p(\mathbf{x} | z, s)$ である。バックワードサンプリングステップでは、コード系列 $\mathbf{S}_n$ を逆順でサンプルする。

$$p(s_{n,d_n} | z_n, d_n, \mathbf{x}_{1:d_n}) \propto \zeta_{n, s_n, d_n} \quad (24)$$

$$p(s_{n,\tau} | z_n, d_n, s_{n,\tau+1:d_n}, \mathbf{x}_{1:d_n}) \propto \zeta_{n, s_n, \tau} \phi_{s_{n,\tau}, s_{n,\tau+1}}^{(z_n)} \quad (25)$$

3.3.4 モデルパラメータのサンプリング

モデルパラメータの更新にギブスサンプリングを用いる。

$$\rho \sim \text{Dirichlet}(\mathbf{a}^\rho + \mathbf{b}^\rho) \quad (26)$$

$$\pi_z \sim \text{Dirichlet}(\mathbf{a}^{\pi_z} + \mathbf{b}^{\pi_z}) \quad (27)$$

$$\psi \sim \text{Dirichlet}(\mathbf{a}^\psi + \mathbf{b}^\psi) \quad (28)$$

$$\mathbf{v} \sim \text{Dirichlet}(\mathbf{a}^\mathbf{v} + \mathbf{b}^\mathbf{v}) \quad (29)$$

$$\phi_s^{(z)} \sim \text{Dirichlet}(\mathbf{a}^\phi + \mathbf{b}^{\phi_s^{(z)}}) \quad (30)$$

$$\Lambda_{z,s}^c \sim \mathcal{W}(\mathbf{W}_{z,s}^c, \nu_{z,s}^c) \quad (31)$$

$$\mu_{z,s}^c | \Lambda_{z,s}^c \sim \mathcal{N}(\mathbf{m}_{z,s}^c, (\beta_{z,s}^c \Lambda_{z,s}^c)^{-1}) \quad (32)$$

$$\Lambda_z^m \sim \mathcal{W}(\mathbf{W}_z^m, \nu_z^m) \quad (33)$$

$$\mu_z^m | \Lambda_z^m \sim \mathcal{N}(\mathbf{m}_z^m, (\beta_z^m \Lambda_z^m)^{-1}) \quad (34)$$

ここで、 $\mathbf{b}^\rho \in \mathbb{R}^K$, $\mathbf{b}^{\pi_z} \in \mathbb{R}^K$, $\mathbf{b}^\psi \in \mathbb{R}^L$, $\mathbf{b}^\mathbf{v} \in \mathbb{R}^K$, および $\mathbf{b}^{\phi_s^{(z)}} \in \mathbb{R}^M$ はサンプルされたデータを数え上げるベクトルである。 b_z^ρ は $z = z_1$ のとき 1, そうでなければ 0 であり, $b_{z'}^{\pi_z}$ はセクション z からセクション z' への遷移回数, b_d^ψ はサンプルされたセクションが継続長 d を持つ回数である。 $b_z^\mathbf{v}$ は $z = z_N$ のとき 1, そうでなければ 0 であり, $b_{s'}^{\phi_s^{(z)}}$ はセクション z におけるコードに関するマルコフ

モデルにおいてコード s からコード s' への遷移回数である。パラメータ $\mathbf{m}_{z,s}^c$, $\beta_{z,s}^c$, $\mathbf{W}_{z,s}^c$ および $\nu_{z,s}^c$ は以下のよう

$$\beta_{z,s}^c = \beta_0^c + N_{z,s}, \quad \nu_{z,s}^c = \nu_0^c + N_{z,s} \quad (35)$$

$$\mathbf{m}_{z,s}^c = \frac{1}{\beta_{z,s}^c} (\beta_0^c \mathbf{m}_0^c + N_{z,s} \bar{\mathbf{x}}_{z,s}^c) \quad (36)$$

$$\begin{aligned} (\mathbf{W}_{z,s}^c)^{-1} &= (\mathbf{W}_0^c)^{-1} + N_{z,s} \mathbf{U}_{z,s}^c \\ &+ \frac{\beta_0^c N_{z,s}}{\beta_0^c + N_{z,s}} (\bar{\mathbf{x}}_{z,s}^c - \mathbf{m}_0^c) (\bar{\mathbf{x}}_{z,s}^c - \mathbf{m}_0^c)^T \end{aligned} \quad (37)$$

ここで、 $N_{z,s}$, $\bar{\mathbf{x}}_{z,s}^c$ および $\mathbf{U}_{z,s}^c$ を以下のように定義する。

$$N_{z,s} = \sum_{t=1}^T \delta_{z_t z} \delta_{s_t s} \quad (38)$$

$$\bar{\mathbf{x}}_{z,s}^c = \frac{1}{N_{z,s}} \sum_{t=1}^T \delta_{z_t z} \delta_{s_t s} \mathbf{x}_t^c \quad (39)$$

$$\mathbf{U}_{z,s}^c = \frac{1}{N_{z,s}} \sum_{t=1}^T \delta_{z_t z} \delta_{s_t s} (\mathbf{x}_t^c - \bar{\mathbf{x}}_{z,s}^c) (\mathbf{x}_t^c - \bar{\mathbf{x}}_{z,s}^c)^T \quad (40)$$

パラメータ $\mathbf{m}_z^m$, β_z^m , $\mathbf{W}_z^m$, ν_z^m も同様に計算する。

3.3.5 ビタビ学習

ギブスサンプリングで得られるサンプルは必ずしも事後分布の局所最適解とは限らないため、パラメータ推定に最終ステップでビタビ学習を適用する。具体的には、フォワードフィルタリング・バックワードサンプリングの代わりにビタビアルゴリズムを適用することで潜在変数を推定し、モデルパラメータの更新にはそれらの事後分布からのサンプリングの代わりに分布の期待値を用いる。ビタビ学習は一般的に近似局所解を求めるのに効果的であることが知られている [29]。

3.3.6 モデルの改良

継続長確率の影響を強めるために重み係数 $w_{\text{dur}} (\geq 1)$ を導入する。具体的には、式 (18) および (19) において確率係数 ψ_d を $(\psi_d)^{w_{\text{dur}}}$ に置き換える。ビタビ学習ステップや 3.4 節で述べる最後の潜在状態推定ステップにおいても同様の置換を行う。また、メルスペクトルとその他の出力確率のバランスをとるために重み係数 w_{label} を導入する。すなわち、式 (9) において出力確率 $\chi_z^s(\mathbf{x}^s)$ を $(\chi_z^s(\mathbf{x}^s))^{w_{\text{label}}}$ に置き換える。

3.4 セクション推定

モデルパラメータ Θ を学習したのち、最大事後確率 (Maximum a Posteriori; MAP) 推定によってセクションラベル系列を推定する。具体的には、事後確率 $p(\mathbf{Z}, \mathbf{D} | \Theta, \mathbf{X})$ をセクションラベル $\mathbf{Z}$ と継続長 $\mathbf{D}$ に関して最大化する。これは、コード系列 $\mathbf{S}$ を積分消去し、セクションラベルに関するマルコフモデルに対して HSMML のビタビアルゴリズム [30] を適用することで求められる。

4. 評価実験

提案手法の性能を調べるため実験を行った。

4.1 実験条件

評価には RWC ポピュラー音楽データベース [31] とその構造アノテーションデータ [32] を用いた。クロマ特徴量は深層特徴量抽出 [33] の結果を、MFCC とメルスペクトルは librosa ライブラリ [34] の出力結果を用いた。ビート情報は madmom ライブラリ [28] の結果を用いた。経験分布 $\mathbf{a}_{\text{emp}}^p$, $\mathbf{a}_{\text{emp}}^\psi$, $\mathbf{a}_{\text{emp}}^\pi$, $\mathbf{a}_{\text{emp}}^v$ は 100 曲に対する leave-one-out 交差検定によって学習した。パラメータの推定では、ギブスサンプリングを 15 回、ビタビ学習を 3 回行った。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。

提案モデルのハイパーパラメータは以下のようにした。 $K = 10$, $L = 40$, $M = 16$, $\sigma = 1$, $w_{\text{dur}} = 4$, $w_{\text{label}} = 0.5$, $\mathbf{a}^p = 1 \cdot \mathbf{a}_{\text{emp}}^p$, $\mathbf{a}^\pi = 1 \cdot \mathbf{a}_{\text{emp}}^\pi$, $\mathbf{a}^\psi = 64 \cdot \mathbf{a}_{\text{emp}}^\psi$, $\mathbf{a}^v = 1 \cdot \mathbf{a}_{\text{emp}}^v$, $\mathbf{a}^\phi = \mathbb{I}$, $\mathbf{m}_0^c = \mathbb{E}[\mathbf{X}^c]$, $\beta_0^c = 64$, $\mathbf{W}_0^c = (\nu_0^c \text{cov}[\mathbf{X}^c])^{-1}$, $\nu_0^c = 512$, $\mathbf{m}_0^m = \mathbb{E}[\mathbf{X}^m]$, $\beta_0^m = 2$, $\mathbf{W}_0^m = (\nu_0^m \text{cov}[\mathbf{X}^m])^{-1}$, $\nu_0^m = 16$ 。ここで、 $\mathbb{I}$ は全ての要素が 1 のベクトルを表す。最初のパラメータ K は [21] で用いられたラベル数に応じて決定した。次の二つのパラメータ L および M は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、多くのセクションの継続長は 40 拍以下であった。あるセクションの継続長が 32 拍（8 小節）であり、各コードの長さが 2 拍であると仮定すると、そのセクションのコード数として 16 が得られる。また、簡単のため $\sigma = 1$ とした。その他の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い決定した。各パラメータはその他のパラメータを固定したうえで、グリッドサーチにより最適化した。各パラメータのさらなる最適化は今後の課題とした。

本研究では、推定結果をセグメンテーション・クラスタリング・ラベリングの観点から評価した。セグメンテーションとクラスタリングの性能は MIREX [35] と同様の評価法により評価した。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 $F_{0.5}$, $F_{3.0}$ [36] により評価した。具体的には、推定境界は $\pm 0.5/3.0$ 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。

クラスタリングの性能は以下のように定義されるペアワイズ F 値 F_{pair} [37] により評価した。推定結果において同じラベルに割り当てられたフレームのペアと、正解データにおいて同様に同じラベルに割り当てられたフレームのペアとを比較した（フレーム長は 100ms とする）。適合率、再現率および F 値は以下のように定義する。

手法	セグメンテーション		クラスタリング	ラベリング
	$F_{0.5}$ (%)	$F_{3.0}$ (%)	F_{pair} (%)	精度 (%)
GS3 [39]	52.3	73.5	54.2	n/a
SUG2 [40]	25.8	73.7	37.3	n/a
FK2 [41]	30.0	65.7	63.4	n/a
Paulus'09 [21]	n/a	63.0	63.7	34.4
カスケード	38.3	63.9	54.9	38.8
ベースライン	30.6	53.5	43.3	39.5
提案モデル	43.3	66.5	54.6	45.3

表 1 比較実験の評価結果

$$P_{\text{pair}} = \frac{|P_E \cap P_A|}{|P_E|}, \quad R_{\text{pair}} = \frac{|P_E \cap P_A|}{|P_A|} \quad (41)$$

$$F_{\text{pair}} = \frac{2P_{\text{pair}}R_{\text{pair}}}{P_{\text{pair}} + R_{\text{pair}}} \quad (42)$$

ここで、 P_E は推定結果において同じラベルに割り当てられたフレームペアの集合、 P_A は正解データにおいて同じラベルに割り当てられたフレームペアの集合を表す。これらの評価値は mir_eval ライブラリ [38] を用いて計算した。ラベリングの性能は [21] と同様にフレーム単位の精度で評価した。これは各フレームにおいて、推定結果と正解データとで割り当てられたラベル同士を比較して計算する。

セグメンテーションとクラスタリングの比較のため、MIREX で公開されている GS3 [39], SUG2 [40], および FK2 [41] を参照した。さらに、全ての観点における比較のため、[21] の結果を引用し、カスケードモデルとベースラインモデルの二つのモデルを用いた。カスケードモデルでは、BLSTM で得られたフレーム単位のラベルを HSMM で得られたクラスターごとに数え、クラスター内で最も多く出現したラベルをそのクラスターにおけるラベルとした。ベースラインモデルは提案モデルと同様の構成を持つが、クロマ特徴量および MFCC を出力せず、ビタビ復号のみを用いて結果を得るモデルである。

4.2 実験結果

表 1 に評価結果を示す。ラベリング精度において、提案手法はラベリングが可能なその他の手法よりも優れた性能を示した。提案モデルは、後処理的に BLSTM のラベリング結果を用いたカスケードモデルと比較して、セグメンテーションとラベリングの性能が高かった。このことは、統一的確率フレームワークによるセグメンテーションとラベリングの同時推論の有効性を示している。また、ビタビ復号のみを用いたベースラインモデルと比較して、提案モデルは全ての評価尺度において優れた性能を示した。このことから、事前分布に基づくベイズ学習の有効性が示された。一方、提案手法はラベリング以外では最高性能を得ることはできなかった。これは、提案手法において同じラベルを持つが継続長が大きく異なる不自然な反復が発生する傾向にあるためだと思われる。一般に、同じラベルを持つ

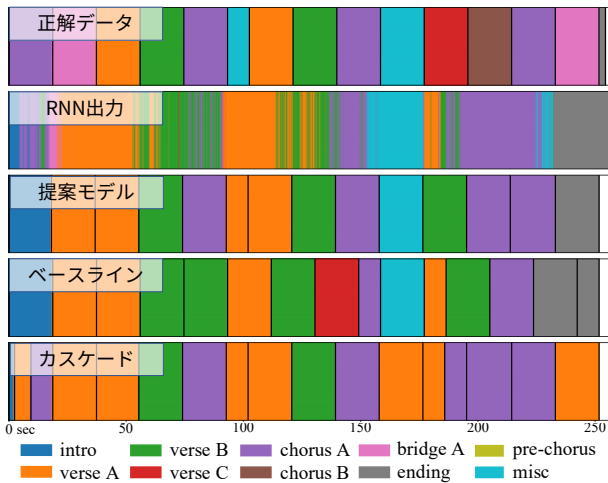


図 3 提案モデル・ベースラインモデル・カスケードモデルの推定結果例 (RWC-MDB-P-2001 No. 25)

セクションはほぼ同じ継続長を持つ。このような制約は、各ラベルに固有の継続確率分布を導入することでモデルに組み込むことができる。

結果例を図 3 に示す。カスケードモデルでは境界位置が正しく推定されながらもラベル付けが誤ったセクションが見られた。これは、HSM みのクラスタリング誤りが伝播したためと考えられる。一方、提案手法では、セグメンテーションとラベリングを同時に行うため、このような誤りが減少した。また、BLSTM での誤りに起因するベースラインモデルの誤りが提案手法の結果では正しく推定されていた。このことから、提案手法は MFCC の同質性とクロマ特徴量の反復性に注目することでこのような誤りを防止する能力があることが示唆された。提案手法は、楽曲の冒頭の「サビ A」を「イントロ」と誤推定していた。このような誤りは、初期確率と出力確率の重みを調節する、あるいは、BLSTM における音楽的に不自然なラベルの切り替えを防ぐために CTC (Connectionist Temporal Classification) 損失関数 [42] などを用いて学習を行うことで避けられる可能性がある。

5. おわりに

本稿では、音楽構造解析のためのベイズ LSTM-HSM ハイブリッドモデルに基づく生成的アプローチを提案した。本モデルはクロマ特徴量、MFCC、メルスペクトルを用いてセクションの重要な性質である、同質性・反復性・規則性を表現する。ベイズ教師なし学習により、セグメンテーションとラベリングが同時に行われる。評価実験の結果により、提案手法が音楽構造解析に有効であることを確かめた。

提案手法は同質性・反復性・規則性を考慮するが、従来研究で強調されているように、新規性を考慮することも重要であるため、この性質をモデルに組み込むことが今後の

課題となっている。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 [43] ことから、より多くの階層を取り扱うように拡張を行うことも重要である [16]。

謝辞 本研究の一部は、JST ACCEL No. JPMJAC1602、JSPS 科研費 No. 16H01744、19K20340 および 19H04137 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] Grohgan, 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] Paulus, J. and Klapuri, A.: Labelling the structural parts of a music piece with Markov models, *Computer Music Modeling and Retrieval (CMMR)*, pp. 166–176 (2008).
- [20] 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).
- [21] 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).
- [22] 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).
- [23] Serrà, J., Müller, M., Grosche, P. and Arcos, J.: Unsupervised music structure annotation by time series structure features and segment similarity, *IEEE Transactions on Multimedia*, Vol. 16, No. 5, pp. 1229–1240 (2014).
- [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] 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).
- [26] Levy, M. and Sandler, M.: New methods in structural segmentation of musical audio, *European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (2006).
- [27] Maddage, N. C., Xu, C., Kankanhalli, M. S. and Shao, X.: Content-based music structure analysis with applications to music semantics understanding, *ACM International Conference on Multimedia (ACMMM)*, pp. 112–119 (2004).
- [28] 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).
- [29] 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).
- [30] Yu, S.-Z.: Hidden Semi-Markov Models, *Artificial Intelligence*, Vol. 174, No. 2, pp. 215–243 (2010).
- [31] 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).
- [32] Goto, M.: AIST Annotation for the RWC Music Database, *International Conference on Music Information Retrieval (ISMIR)*, pp. 359–360 (2006).
- [33] 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).
- [34] 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).
- [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] Turnbull, D., Lanckriet, G., Pampalk, E. and Goto, M.: A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting, *International Conference on Music Information Retrieval (ISMIR)*, pp. 51–54 (2007).
- [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] Grill, T. and Schlüter, J.: Structural segmentation with convolutional neural networks MIREX submission, *MIREX* (2015).
- [40] Schlüter, J., Ullrich, K. and Grill, T.: Structural segmentation with convolutional neural networks MIREX submission, *MIREX* (2014).
- [41] Kaiser, F. and Peeters, G.: Music structural segmentation task: IrcamStructure submission, *MIREX* (2013).
- [42] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *International Conference on Machine Learning (ICML)*, pp. 369–376 (2006).
- [43] Lerdahl, F. and Jackendoff, R.: *A Generative Theory of Tonal Music*, MIT Press (1983).