Vol.2022-MUS-135 No.6
2022/9/15

TCN-HSMM ハイブリッドモデルに基づくビート・ダウンビート推定

大山 偉永[1,a)]　中村 栄太[1,2,b)]　吉井 和佳[1,3,c)]

概要：本稿では，可変拍子を含む楽曲を適切に取り扱うことができる，深層生成モデルに基づくビート・ダウンビート推定手法を提案する。ポピュラー音楽の多くが，拍子が一時的に変化する可変拍子を含んでおり，可変拍子を考慮したビート・ダウンビート推定は重要な課題となっている．従来の標準的なビート・ダウンビート推定は二段階で構成されており，例えば，時間畳み込みネットワーク (TCN) を用いて各時刻におけるビート・ダウンビートの事後分布を推定した後，周期性を考慮できる隠れマルコフモデル (HMM) を用いてビート・ダウンビート時刻の検出を行う。しかし，HMM が一意な拍子を仮定しているため，可変拍子の楽曲に対応できず不自然な推定結果となることがあった．そこで，本研究では，HMM に拍子の生成モデルを導入し，拍子・ビート・ダウンビートに関する潜在表現から観測の音響特徴量を生成する過程を表す隠れセミマルコフモデル (HSMM) を提案する．音響特徴量の観測確率は TCN が推定したビート・ダウンビートの事後確率から計算し，ビタビアルゴリズムにより最尤ビート・ダウンビート系列を推定する．実験により提案手法の有効性を確認する．

1. はじめに

ビート・ダウンビート推定は自動ピアノ採譜 [1]，音楽構造解析 [2]，ドラム採譜 [3, 4] など音楽情報処理分野のタスクにおいて重要な基礎技術となっている．ここで，本稿の対象とするポピュラー音楽の多くは楽曲内で拍子が変化する可変拍子を含んでおり，このような性質を考慮したビート・ダウンビート推定は重要な課題となっている．実際に本稿で使用するポピュラー音楽 2436 曲のうち約 25% の 602 曲が拍子変化を含んでいる．さらに，可変拍子の種類として，セクション間での拍子変化といった拍子変化の間隔が大きいものや，楽曲全体を通じて 4/4 拍子であるがサビ前に 2/4 拍子の小節が一つ挿入されるといった，局所的に拍子が変化するものもあり，両者を適切にモデル化し推定することが重要である．

近年のビート・ダウンビート推定手法 [5–11] の多くは，(1)DNN を用いてフレームごとのビート・ダウンビート存在確率を推定 (2)DNN の出力を隠れマルコフモデル (HMM) に入力してビート・ダウンビート時刻を推定，の二段階で構成されている．具体的には，時間畳み込みネットワーク (TCN) [6] や Transformer [11] によってビート・ダウンビートの事後確率を推定し，拍節構造の周期性に関する知識を内包した HMM の一種である dynamic Bayesian network(DBN) [12–14] によって各時刻の検出を行う．ただし，これらの手法で用いられている DBN はテンポと小節内の位置に関する状態のみ持ち，楽曲を通して一定の拍子を仮定するため，可変拍子を含む楽曲に対して不自然な推定結果となることがあった．

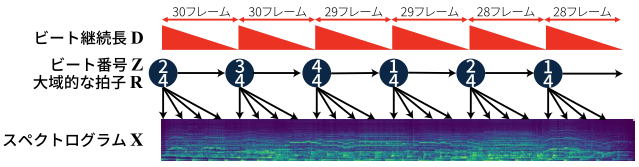

図 1　大域的な拍子系列を導入した HSMM

本稿では，既存の DBN の潜在状態に加え，新たに拍子の生成モデルを導入し，潜在状態から音響特徴量を観測として生成する隠れセミマルコフモデル (HSMM) を提案する．また，セクション単位の拍子変化と局所的な拍子変化を同時に捉えるため，拍子の生成モデルに加え，局所的な拍子変化に依存しない大域的な拍子の生成モデルを導入する．音響特徴量の観測確率は TCN が推定したビート・ダウンビートの存在確率から計算を行う．ビタビアルゴリズムにより最尤状態系列を推定する．可変拍子を含む楽曲に対して，拍子変化を仮定しない従来手法 [8] と提案手法を比較し，拍子のモデル化の効果を確認する．

1　京都大学 大学院情報学研究科
2　京都大学 白眉センター
3　科学技術振興機構 戦略的創造研究推進事業（さきがけ）
a)　ooyama@sap.ist.i.kyoto-u.ac.jp
b)　enakamura@sap.ist.i.kyoto-u.ac.jp
c)　yoshii@i.kyoto-u.ac.jp

2. 関連研究

古典的なビート推定手法では，(1) 音響信号からビートに関する特徴量を抽出，(2) 特徴量からビートの周期を推定，(3) 特徴量と周期からビートの位相を推定，の 3 ステップで構成されていた [15, 16]．ステップ (1) の特徴量としては，オンセット確率系列 [15–17] やコードの変化 [18] などが用いられていたが，近年の DNN の発展により，DNN を用いて直接各時刻のビート存在確率を推定する手法が主流となった．DNN を用いた初期の手法の多くが，長短期記憶 (LSTM) ネットワークを使用したが [12, 19, 20]，学習に長時間を要することから，時間畳み込みネットワーク (TCN) [21] を用いたビート推定が提案され [6]，ビート推定における現在の標準的なアプローチの一部となっている．また Transformer モデルの隆盛により，ビート推定でも Transformer を組み込んだ DNN モデルが提案されており，高い精度を示している [11]．ステップ (2) と (3) に関しては，ステップ (1) で推定された特徴量を入力として，テンポとビートの位相を同時にモデル化する dynamic Bayesian network(DBN) [12] が提案され，近年のビート推定の標準的な後処理手法となっている．

ダウンビート推定は，事前に推定されたビート時刻を用いる手法 [20, 22, 23] とビート推定と同時に行う手法 [5, 7, 8, 11] がそれぞれ提案されている．後者の手法では，一つの DNN をビート推定とダウンビート推定の 2 つのタスクで共有するマルチタスク学習により高い相乗効果が示されており，ビートと同時にダウンビートを推定する手法が現在の標準的な枠組みとなっている．またテンポも同時にマルチタスク学習することでより周期的なビート・ダウンビート推定が可能になったことも示されている [7, 8]．これらの手法では，後処理で用いられる DBN もダウンビートをビート・テンポと同時にモデル化するように拡張され，DNN で推定されたビートとダウンビートそれぞれの存在確率が入力される．しかし，この DBN は各曲に単一の拍子のみを仮定してダウンビートをモデル化しており，日本のポピュラー音楽のように拍子の変化の多いジャンルには適さず，拍子の変化を捉えるモデルの導入が重要な課題である．

3. 提案法

本章では，可変拍子に対応可能な三つの TCN-HSMM ハイブリッドモデルを説明する．3.1 節 で各モデル共通の問題設定，3.2 節 でベースラインモデル，3.3 節 で拍子の生成モデルを導入したモデル，3.4 節 で局所的な拍子の代わりに大域的な拍子の生成モデルを導入したモデル，3.5 節 で各モデルのパラメータの学習・推論方法について説明する．

3.1 問題設定

以下で，本稿が取り扱う問題を定義する．ここで T は入力のフレーム数，N は曲中のビート数とする．

入力：音楽音響信号から得られた，ログスペクトログラム $\mathbf{X} \triangleq \{\mathbf{x}_t\}_{t=1}^{T}$

出力：ビート番号系列 $\mathbf{Z} \triangleq \{z_n\}_{n=1}^{N} (z_n \in \{1, 2, 3, 4\})$ と各ビートの時刻系列 $\mathbf{T} \triangleq \{t_n\}_{n=1}^{N}$

ここで，ビート番号は各ビートが小節内で何番目のビートかを表す値であり，本稿ではビート番号が z' の状態を z' 拍目と表記する．また各ビートの継続長 (局所テンポ) 系列を $\mathbf{D} \triangleq \{d_n\}_{n=1}^{N} (d_n \in \{d_{\min}, d_{\min}+1, ..., d_{\max}\})$ とする．ここで $d_{\min}$ と $d_{\max}$ はそれぞれ継続長の最小値と最大値を表し，$\sum d_n = T$ が成立する．

3.2 ベースラインモデル

本節では，ビート番号に関する生成モデルとログスペクトログラムを観測する音響モデルで構成される TCN-HSMM ハイブリッドモデルについて説明する．

3.2.1 モデル定式化

潜在系列としてビート番号系列 $\mathbf{Z}$，観測としてログスペクトログラム $\mathbf{X}$ をもつ HSMM を定式化する．このモデルの同時確率は以下で与える．

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

ここで，$p(\mathbf{X}|\mathbf{Z})$ は観測特徴量 $\mathbf{X}$ に関する音響モデル，$p(\mathbf{D})$ は継続長系列 $\mathbf{D}$ に関する全遷移型セミマルコフモデル，$p(\mathbf{Z})$ はビート番号系列 $\mathbf{Z}$ に関する全遷移型セミマルコフモデルである．

3.2.2 継続長に関するマルコフ連鎖

式 (1) 中の全遷移型セミマルコフモデル $p(\mathbf{D})$ は，以下のように継続長系列 $\mathbf{D}$ の生成過程を表現する．

$$p(\mathbf{D}) = p(d_1) \prod_{n=2}^{N} p(d_n|d_{n-1}) \tag{2}$$

各項は以下で与える．

$$p(d_1) = \frac{1}{d_{\max} - d_{\min} + 1} \tag{3}$$

$$p(d_n|d_{n-1}) = \exp\left(-\lambda \times \left|\frac{d_n}{d_{n-1}} - 1\right|\right) \tag{4}$$

ここで，式 (4) の遷移確率は従来手法 [13] と同様のものを用いる．λ はテンポの変わりやすさに関するパラメータであり，値が大きければテンポの変化は小さくなる．なお，本節の継続長に関するモデルは後述の手法でも同様のため，以下では説明を省略する．

3.2.3 ビート番号に関するマルコフ連鎖

式 (1) 中の全遷移型セミマルコフモデル $p(\mathbf{Z})$ は，以下のようにビート番号系列 $\mathbf{Z}$ の生成過程を表現する．

$$p(\mathbf{Z}) = p(z_1)\prod_{n=2}^{N} p(z_n|z_{n-1}) \tag{5}$$

それぞれの項は以下で与える．

$$p(z_1) = \frac{1}{|z_1|} \tag{6}$$

$$p(z_n|z_{n-1}) = A_{z_{n-1}z_n} \tag{7}$$

$$A = \begin{bmatrix} \frac{1}{101} & \frac{100}{101} & 0 & 0 \\ \frac{1}{101} & 0 & \frac{100}{101} & 0 \\ \frac{1}{11} & 0 & 0 & \frac{10}{11} \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{8}$$

ここで $|y|$ は y が取りうる値の数を表す．A はビート番号系列 $\mathbf{Z}$ の遷移確率行列であり，i 行 j 列の値 A_{ij} はビート番号が i から j に遷移する確率を表す．各行の一列目の値を正の値に設定することで，任意のビート番号から一拍目への遷移を可能にしている．一般にポピュラー音楽の多くが四拍子で構成されているため，遷移確率行列 A はビート番号が $1 \to 2 \to 3 \to 4$ のような遷移を行う確率が高くなるように設定する．また一拍子・二拍子の曲の数に対して三拍子の曲の数のほうが多いため，三拍目から一拍目に遷移する確率 A_{31} は一・二拍目から一拍目に遷移する確率 A_{11} と A_{21} に比べて大きく設定する．

遷移確率行列 A は先述の通り，四拍子以外の曲に対しても四拍目まで遷移しやすい性質 ($A_{23} > A_{21}$, $A_{34} > A_{31}$) を持つため，本モデルは四拍子の曲以外に対して不適であり，各拍子に関して適当な遷移確率を定める必要がある．

3.2.4 音響特徴量の出力

式 (1) 中の音響モデル $p(\mathbf{X}|\mathbf{Z})$ は，以下のようにログスペクトログラム $\mathbf{X}$ と適当な系列 $\mathbf{Z}'$ の同時確率分布を周辺化することで計算できる．

$$p(\mathbf{X}|\mathbf{Z}) = \sum_{\mathbf{Z}'} p(\mathbf{X}, \mathbf{Z}'|\mathbf{Z}) = \sum_{\mathbf{Z}'} p(\mathbf{X}|\mathbf{Z}')p(\mathbf{Z}'|\mathbf{Z}) \tag{9}$$

ここで系列 $\mathbf{Z}'$ として，ビートがあるフレームにのみビート番号 z_n が現れ，それ以外のフレームでは 0 となる長さ T のアクティベーション系列 $\hat{\mathbf{Z}} \triangleq \{\hat{z}_t\}_{t=1}^{T}$ ($\hat{z}_t \in \{0,1,2,3,4\}$) を考える．このとき，系列 $\hat{\mathbf{Z}}$ をビートごとに分割し $\hat{\mathbf{Z}} = \{\hat{\mathbf{Z}}_n\}_{n=1}^{N}$, $\hat{\mathbf{Z}}_n = \{\hat{z}_{n\tau}\}_{\tau=1}^{d_n}$ とすると，定義より以下を満たす．

$$\hat{z}_{n\tau} = \begin{cases} z_n & (\tau = 1) \\ 0 & (\text{otherwise}) \end{cases} \tag{10}$$

このとき，ビート番号系列 $\mathbf{Z}$ に対応するアクティベーション系列 $\hat{\mathbf{Z}}$ は一意に定まるため，式 (9) の音響モデルは以下のように書ける．

$$p(\mathbf{Z}'|\mathbf{Z}) = \delta_{\mathbf{Z}'\hat{\mathbf{Z}}} \tag{11}$$

$$p(\mathbf{X}|\mathbf{Z}) = \sum_{\mathbf{Z}'} p(\mathbf{X}|\mathbf{Z}')\delta_{\mathbf{Z}'\hat{\mathbf{Z}}} = p(\mathbf{X}|\hat{\mathbf{Z}}) \tag{12}$$

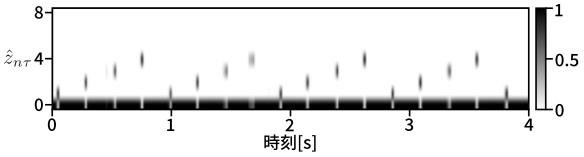

図 2 四拍子の曲に対して TCN で推定した $p(\hat{\mathbf{Z}}|\mathbf{X})$ の例.

ここで δ_{ij} は $i = j$ のときに 1，そうでなければ 0 となるクロネッカーのデルタを表す．式 (12) より，音響モデル $p(\mathbf{X}|\mathbf{Z})$ はアクティベーション系列 $\hat{\mathbf{Z}}$ からログスペクトログラム $\mathbf{X}$ を観測する確率 $p(\mathbf{X}|\hat{\mathbf{Z}})$ と同値であり，以下で与える．

$$p(\mathbf{X}|\hat{\mathbf{Z}}) = \prod_{t=1}^{T} p(\mathbf{x}_t|\hat{z}_t) \tag{13}$$

$$p(\mathbf{x}_t|\hat{z}_t) \propto \frac{p(\hat{z}_t|\mathbf{x}_t)}{p(\hat{z}_t)} \tag{14}$$

式 (14) の $p(\hat{z}_t)$ はアクティベーション値 $\hat{z}_t$ のユニグラム確率であり，以下で与える．

$$p(\hat{z}_t) = \begin{cases} \psi & (\hat{z}_t = 0) \\ \frac{1-\psi}{|\hat{z}_t|-1} & (\text{otherwise}) \end{cases} \tag{15}$$

ここで，ψ はアクテベーション系列 $\hat{\mathbf{Z}}$ 内で $\hat{z}_t = 0$ となる確率を表すパラメータである．また，式 (14) の $p(\hat{z}_t|\mathbf{x}_t)$ はログスペクトログラム $\mathbf{X}$ を入力として学習された DNN から得る（図 2）．DNN のアーキテクチャは基本的には TCN を用いた従来手法 [8] と同様のものを用いるが，本研究では TCN に一つの線形層を接続し $T \times K$ 次元の出力 π_{tk} を得る．ここで，K は学習データにおける一小節内の最大ビート数，π_{tk} はフレーム t に k 番目のビートが存在する確率を表す．本研究ではデータセット内に一小節に 8 ビートある楽曲が存在するため，図 2 のように $K = 8$ とした．また本稿で扱うビート番号 z_n の最大値は 4 であるため，以下のように出力確率 π_{tk} の $k \leq 4$ の箇所を正規化して観測確率 $p(\hat{z}_t|\mathbf{x}_t)$ を計算する．

$$p(\hat{z}_t|\mathbf{x}_t) = \frac{\pi_{t\hat{z}_t}}{\sum_{k=1}^{4}\pi_{tk}} \tag{16}$$

3.2.2 節の継続長系列 $\mathbf{D}$ に関するモデルと同様に，本節の音響モデルは後述の手法でも同様のため，以下では説明を省略する．

3.3 拍子の生成モデルを導入したモデル

本節では，HSMM に拍子の生成モデルを導入し，各拍子に関して独立したビート番号の遷移確率を定義することでベースラインモデルの問題を解決する．

3.3.1 モデル定式化

ベースラインモデルに拍子系列 $\mathbf{S} = \{s_n\}_{n=1}^{N} (s_n \in \{1,2,3,4\})$ に関する生成モデルを導入し，同時確率を以下で定義する．

$$p(\mathbf{S},\mathbf{D},\mathbf{Z},\mathbf{X}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{S},\mathbf{Z})p(\mathbf{D}) \tag{17}$$

ここで，$p(\mathbf{X}|\mathbf{Z})$ と $p(\mathbf{D})$ はベースラインモデルと同様である．$p(\mathbf{S},\mathbf{Z})$ は拍子系列 $\mathbf{S}$ およびビート番号系列 $\mathbf{Z}$ に関する全遷移型セミマルコフモデルである．また本モデルの生成モデル $p(\mathbf{S},\mathbf{Z})$ は従来手法 [20] と同様のものである．

3.3.2 拍子とビート番号に関するマルコフ連鎖

式 (17) 中の全遷移型セミマルコフモデル $p(\mathbf{S},\mathbf{Z})$ は，以下のように拍子系列 $\mathbf{S}$ とビート番号系列 $\mathbf{Z}$ の生成過程を表現する．

$$p(\mathbf{S},\mathbf{Z}) = p(s_1)p(z_1)\prod_{n=2}^{N} p(s_n,z_n|s_{n-1},z_{n-1}) \tag{18}$$

それぞれの項は以下で与える．

$$p(s_1) = \frac{1}{|s_1|},\ p(z_1) = \frac{1}{|z_1|} \tag{19}$$

$$p(s_n,z_n|s_{n-1},z_{n-1}) = p(s_n|z_n,s_{n-1},z_{n-1}) \times p(z_n|s_{n-1},z_{n-1}) \tag{20}$$

ここで，式 (20) の第一項は拍子の遷移確率，第二項はビート番号の遷移確率を表す．拍子の遷移確率は以下で与える．

$$p(s_n|z_n,s_{n-1},z_{n-1}) = \begin{cases} \pi_{s_{n-1}s_n} & (z_n = 1) \\ \delta_{s_{n-1}s_n} & (\text{otherwise}) \end{cases} \tag{21}$$

$$\pi_{s_{n-1}s_n} = \begin{cases} 1 - c\epsilon_s & (s_{n-1} = s_n) \\ \epsilon_s & (\text{otherwise}) \end{cases} \tag{22}$$

ここで，$c = |s_{n-1}| - 1$ であり，ϵ_s は拍子が前の状態とは異なる値に遷移する確率を表す微小パラメータを表す．拍子の遷移は，ビート番号 $z_n = 1$ のときのみ，つまりビートが一拍目に遷移するときのみ許されており，式 (21) の $z_n = 1$ の場合で表される．式 (20) のビート番号の遷移確率 $p(z_n|s_{n-1},z_{n-1})$ は以下で与える．

$$p(z_n|s_{n-1},z_{n-1}) = \begin{cases} \frac{1}{c} & (s_{n-1} = 1) \\ 1 & (z_n = (z_{n-1} \bmod s_{n-1}) + 1) \\ 0 & (\text{otherwise}) \end{cases} \tag{23}$$

式 (23) は一拍子 $(s_{n-1} = 1)$ の時を除いて，ビート番号が拍子 s_{n-1} に従って一意な遷移をすることを表しており，直前のビート番号 z_{n-1} が拍子 s_{n-1} と等しいときは一拍目に戻り $(z_n = 1)$，そうでなければ次の拍に遷移する $(z_n = z_{n-1} + 1)$ ことを表す．図 3 に本モデルの状態遷移図を示す．

本モデルでは，拍子の確率変数を導入することで，ベースラインモデルの問題を解決し，各楽曲の拍子に適したビート番号の遷移が可能となる．ただし，ポピュラー音楽においては，楽曲中で拍子が完全に変化する場合に比べ，

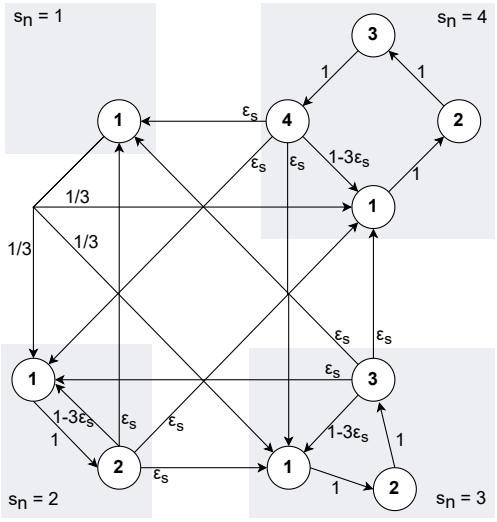

図 3 状態遷移図．丸の中の値がビート番号 z_n，辺に与えられている値が遷移確率を表す．

サビ直前などで一小節のみ別拍子が挟まるといった，局所的な拍子の変化が多い．しかし，本モデルでは別拍子への遷移確率は低く設定され，別拍子への遷移後も遷移前の拍子状態に関する記憶は保持されないため，このような局所的な拍子変化への対応が難しい性質を持つ．

3.4 大域的な拍子の生成モデルを導入した手法

本節では，前節の拍子の生成モデルの代替として，大域的な拍子の生成モデルを導入したモデルを提案する．ここで，大域的な拍子とは一時的な拍子の変化に依存しない各時点での支配的な拍子を表す．具体的には，図 1 のように実際の拍子が二拍子の小節においても，前後の小節が四拍子で一時的な拍子の変化とみなされる場合は，大域的な拍子は四拍子として継続される．

3.4.1 モデル定式化

図 1 のように，3.3 節の拍子系列 $\mathbf{S}$ の代わりに，大域的な拍子系列 $\mathbf{R} = \{r_n\}_{n=1}^{N}(r_n \in \{3,4\})$ を導入し，このモデルの同時確率を以下で定義する．

$$p(\mathbf{R},\mathbf{D},\mathbf{Z},\mathbf{X}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{R},\mathbf{Z})p(\mathbf{D}) \tag{24}$$

ここで，$p(\mathbf{X}|\mathbf{Z})$ と $p(\mathbf{D})$ はベースラインモデルと同様である．$p(\mathbf{R},\mathbf{Z})$ は大域的な拍子系列 $\mathbf{R}$ およびビート番号系列 $\mathbf{Z}$ に関する全遷移型セミマルコフモデルである．以降，本節では大域的な拍子系列を単に拍子系列と記す．

3.4.2 拍子とビート番号に関するマルコフ連鎖

式 (24) 中の全遷移型セミマルコフモデル $p(\mathbf{R},\mathbf{Z})$ は，以下のように拍子系列 $\mathbf{R}$ とビート番号系列 $\mathbf{Z}$ の生成過程を表現する．

$$p(\mathbf{R},\mathbf{Z}) = p(r_1)p(z_1)\prod_{n=2}^{N} p(r_n,z_n|r_{n-1},z_{n-1}) \tag{25}$$

それぞれの項は以下で与える．

表 1　可変拍子を含む楽曲に対してビート・ダウンビート推定を行った評価結果.

	ビート					ダウンビート				
手法	F 値	CMLc	CMLt	AMLc	AMLt	F 値	CMLc	CMLt	AMLc	AMLt
ベースラインモデル	**89.3**	**73.6**	**80.3**	**79.1**	**87.5**	80.0	51.5	**73.2**	55.4	79.5
拍子モデル	89.1	**73.6**	80.2	**79.1**	87.3	79.6	47.3	68.3	52.2	76.5
大域的な拍子モデル	89.0	73.3	79.9	78.9	87.1	**80.4**	51.9	**73.2**	56.5	**80.6**
従来手法 [8]	88.5	61.9	79.0	68.1	87.0	76.8	**53.7**	69.9	**60.4**	78.7

$$p(r_1) = \frac{1}{|r_1|},\ p(z_1) = \frac{1}{|z_1|} \tag{26}$$

$$p(r_n, z_n | r_{n-1}, z_{n-1}) = p(r_n | z_n, r_{n-1}, z_{n-1}) \times p(z_n | r_{n-1}, z_{n-1}) \tag{27}$$

ここで，式 (27) の第一項は拍子の遷移確率は式 (21) と同様のものを用いる．本モデルにおける拍子遷移に関するパラメータは ϵ_r とする．第二項のビート番号の遷移確率 $p(z_n|r_{n-1}, z_{n-1})$ は以下で与える．

$$p(z_n | r_{n-1}, z_{n-1}) = A^{(r_{n-1})}_{z_{n-1} z_n} \tag{28}$$

ここで，$A^{(r_{n-1})}$ はビート番号に関する遷移確率行列であり，r_{n-1} に関してそれぞれ以下で与える．

$$A^{(3)} = \begin{bmatrix} \epsilon_b & 1-\epsilon_b & 0 & 0 \\ \epsilon_b & 0 & 1-\epsilon_b & 0 \\ 1-\epsilon_b & 0 & 0 & \epsilon_b \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{29}$$

$$A^{(4)} = \begin{bmatrix} \epsilon_b & 1-\epsilon_b & 0 & 0 \\ \epsilon_b & 0 & 1-\epsilon_b & 0 \\ \epsilon_b & 0 & 0 & 1-\epsilon_b \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{30}$$

ここで ϵ_b は一時的に拍子が変化する確率を表す微小パラメータを表す．遷移確率行列 $A^{(r_{n-1})}$ は式 (23) を遷移確率行列として各拍子 s_{n-1} に関して展開し，任意のビート番号から一拍目への遷移も可能にした行列と解釈される．これは，式 (8) を各拍子に関して定義した行列と捉えることも可能である．したがって，本モデルは拍子の生成モデルを導入したモデルのビート番号に関する遷移確率としてベースラインモデルの遷移確率を採用したモデルと解釈でき，これにより大域的な拍子の変化と局所的な拍子の変化を同時にモデル化可能となる．

3.5　モデルパラメータの学習と HSMM の推論

各モデルの拍子・ビート番号の遷移確率に関するパラメータはグリッドサーチを行い検証データに対して一番良い精度を示したものを使用する．また遷移確率行列をデータセットから最尤推定により学習を行うこともできるが，グリッドサーチによる結果と有意な差は見られなかった．同様に Viterbi 学習も行ったが効果は見られなかった．最尤ビート番号系列は Viterbi アルゴリズムにより推定する．

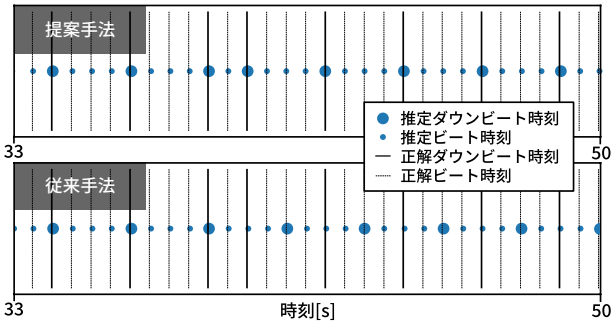

図 4　あいみょん「葵」[*2]に対する提案手法 (大域的な拍子モデル) と従来手法 [8] の推定結果例．基本的に四拍子で，一時的に二拍子が挟まる例.

4.　評価実験

4.1　実験設定

DNN のパラメータや学習方法は，[9] と同様のものを用いた．またデータオーグメンテーションは [8] と同様の手法を用いた．DNN の学習にはビートとダウンビートの正解ラベルが存在する，Ballroom [24,25], Hainsworth [26], Beatles [27], HJDB [28] を用いた．評価には YouTube より取得した日本のポピュラー音楽 2436 曲を使用した．HSMM の最大継続長・最小継続長はそれぞれ $d_{\max} = 120, d_{\min} = 25$ とした．継続長の遷移確率に関するパラメータ λ は [8] と同様に $\lambda = 100$ とした．観測確率 $p(\mathbf{X}|\hat{\mathbf{Z}})$ の事前分布に関するパラメータ ψ は $\psi = 0.96$ とした．

評価指標は [8] と同様に，ビート・ダウンビート推定には F 値，CMLc, CMLt，AMLc, AMLt を用いた．F 値は正解フレームの前後 ± 70 ms の範囲を正解とし，適合率と再現率から計算する．また，あるビートを正解とする条件を「推定ビートの局所テンポと位相が正解の ±17.5% 以内にあること」と定義すると，CMLc は楽曲中の全ビートの中で正解条件を連続で満たすビートの最大長の割合，CMLt は全ビート数の中で正解条件を満たしたビート数の割合として計算する．正解ビート系列に対する CMLc と CMLt の評価値とともに，正解ビート系列の倍テンポ・半テンポに対応するビート系列に対しても CMLc と CMLt の評価を行い，三つの系列に対する評価値の中で最も高いものをそれぞれ AMLc と AMLt とする．評価実験として，本稿の三つの TCN-HSMM ハイブリッドモデルと従来法 [8] の

[*2] https://youtu.be/B6EKkD2QugM?t=33

表 2 楽曲中の拍子変化の回数ごとの大域的な拍子モデルの評価結果．以下の各行は上から順に楽曲中の拍子変化の回数が 0 回 (1834 曲)，1 回以上 5 回未満 (455 曲)，5 回以上 10 回未満 (111 曲)，10 回以上 (36 曲) にそれぞれ対応する．括弧内の値は従来手法 [8] の評価値を表し，下線は二手法の内で精度の高いものを示す．

	ビート					ダウンビート				
	F 値	CMLc	CMLt	AMLc	AMLt	F 値	CMLc	CMLt	AMLc	AMLt
1	92.8(91.9)	82.3(79.5)	86.8(84.7)	86.4(85.5)	92.1(92.1)	86.4(87.6)	78.0(82.7)	83.4(83.9)	82.0(90.1)	88.2(91.7)
2	89.3(88.5)	73.8(62.7)	80.6(79.4)	78.7(68.7)	87.1(87.1)	81.6(79.4)	57.0(59.4)	75.4(73.0)	61.9(66.9)	82.2(82.2)
3	88.4(88.2)	73.2(59.2)	79.4(78.3)	80.6(65.9)	87.6(86.7)	77.6(70.3)	38.2(39.4)	68.4(62.1)	42.1(43.8)	77.7(69.9)
4	86.9(89.3)	66.7(60.2)	72.4(77.5)	75.1(68.1)	85.2(85.9)	73.9(64.8)	29.7(26.1)	59.3(54.6)	33.3(29.0)	69.2(60.7)

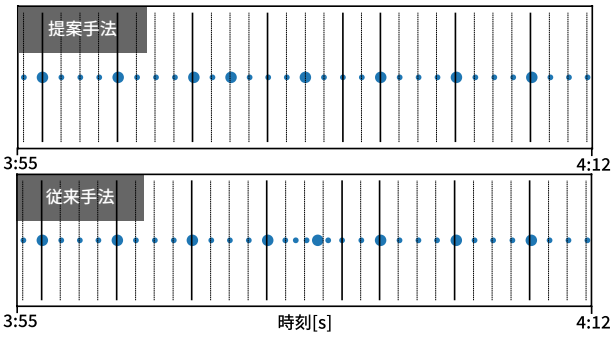

図 5 レミオロメン「太陽の下」[*4]に対する従来手法 [8] の推定結果例．基本的に四拍子で，一時的に二拍子が挟まる例．

比較を可変拍子が含まれる楽曲 602 曲に対して行い，提案手法の可変拍子を含む楽曲に対する効果を確認する．次に，楽曲内の拍子の変化回数に応じた評価結果に対して考察を行う．

4.2 実験結果

図 4 に可変拍子を含む楽曲に対する推定結果例を示す．提案手法では一時的に挟まる二拍子の小節に対して正しく推定が行われ，その後も正しくダウンビートが推定されているが，従来手法では拍子の変化はモデル化されていないため，二拍子の小節以後ダウンビートが裏拍箇所に誤って推定されている．

可変拍子を含む楽曲に対してビート・ダウンビート推定を行った評価結果を表 1 に示す．ビート推定結果は従来手法 [8] に比べ，本稿で説明を行った可変拍子に対応した三手法が各尺度で高い精度を示した．特に CMLc と AMLc で約 10 ポイントの精度改善が確認された．従来手法では，図 5 のように，モデルの尤度を高めるため一時的にビート間隔を縮めた推定結果例が散見されるが，提案手法では可変拍子区間においても基本的にビートは一定間隔で推定されるため，各楽曲の長さに対して連続で正しい推定が行われた区間の割合を表す CMLc と AMLc で有意な差が生まれたと考えられる．このように，可変拍子のモデル化がビート推定の安定性を向上させる効果が確認された．ダウンビート推定に関しては，可変拍子に対応した三手法は拍子変化に対する柔軟な追従性により F 値・CMLt・AMLt で高い精度を示した．一方，CMLc と AMLt では，ビート推定とは逆に従来手法が上回る結果となった．これは，拍子変化のない区間においても提案手法が誤って拍子の変化を捉えてしまったためだと考えられる．三手法の比較に関しては，ビート推定においては有意な差は認められなかったが，ダウンビート推定においては大域的な拍子の生成モデルを導入したモデルが高い精度を示した．本稿の評価データには楽曲全体で三拍子の曲が含まれなかったため，四拍子の楽曲に合わせて遷移確率行列が設定されたベースライン手法は高い精度を示したと考えれ

楽曲中の拍子変化の回数に応じたクラスごとの評価結果を表 2 に示す．ビートに関しては，従来手法は拍子変化が一回以上ある各クラス間で有意な差はないが，提案手法は拍子変化の回数が多いクラスほど精度が悪化する傾向がある．提案手法と従来手法では DNN の出力形式が違うため，拍子の多いクラスダウンビートは拍子変化の回数が多いクラスほど従来手法と比較して優位になる傾向があるが，精度はいずれも低くなっている．これは，図 5 のように，提案手法は可変拍子を捉えることにより一定間隔のビート推定が可能な一方で，拍子が変化する小節の推定は高い精度で行えないことが原因である．ただし，拍子変化位置の正確な推定は DNN の出力への依存が大きく DNN の推定の改善や他のモデル化が必要だと考えられる．

5. おわりに

本稿では，可変拍子を含む楽曲に対するビート・ダウンビート推定を生成モデルとして提案した．既存の拍子の生成モデルに対して大域的な拍子の生成モデルを導入することでポピュラー音楽に散見される楽曲中での局所的な拍子の変化を捉えることのできる手法を提案した．評価実験より，提案手法がポピュラー音楽に対してより安定した推定を行えることを確認した．

今後の課題として，提案手法は従来手法と比べて潜在状態が多く推定により時間がかかるため，枝刈りなどにより不必要な状態に対する計算を省略する方法を検討する．また，本稿では DNN を用いて各ビート番号に対するビート存在確率を推定し HSMM の観測確率として用いたが，今後は拍子の推定も同時に行い HSMM の観測確率として使

[*4] https://youtu.be/dtkSHTHMv7w?t=235

用することも検討する．また楽曲の A メロやサビなど同じセクションでは同様の拍節構造が繰り返される傾向があるため，音楽構造解析手法からセクションの生成モデルを導入することを検討する．

謝辞 本研究の一部は，JSPS 科研費 No. 19H04137，No. 20K21813 および JST さきがけ No. JPMJPR20CB の支援を受けた．

参考文献

[1] Nishikimi, R., Nakamura, E., Goto, M., Itoyama, K. and Yoshii, K.: Bayesian Singing Transcription Based on a Hierarchical Generative Model of Keys, Musical Notes, and F0 Trajectories, *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 1678–1691 (2020).

[2] Shibata, G., Nishikimi, R. and Yoshii, K.: Music Structure Analysis Based on an LSTM-HSMM Hybrid Model, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 15–22 (2020).

[3] Ishizuka, R., Nishikimi, R., Nakamura, E. and Yoshii, K.: Tatum-Level Drum Transcription Based on a Convolutional Recurrent Neural Network with Language Model-Based Regularized Training, *APSIPA*, pp. 359–364 (2020).

[4] Ishizuka, R., Nishikimi, R. and Yoshii, K.: Global Structure-Aware Drum Transcription Based on Self-Attention Mechanisms, *arXiv* (2021).

[5] 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).

[6] MatthewDavies, E. and Böck, S.: Temporal Convolutional Networks for Musical Audio Beat Tracking, *European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (2019).

[7] Böck, S., Davies, M. E. and Knees, P.: Multi-Task Learning of Tempo and Beat: Learning One to Improve the Other, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 486–493 (2019).

[8] Böck, S. and Davies, M. E.: Deconstruct, Analyse, Reconstruct: How To Improve Tempo, Beat, and Downbeat Estimation, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 574–582 (2020).

[9] Oyama, T., Ishizuka, R. and Yoshii, K.: Phase-Aware Joint Beat and Downbeat Estimation Based on Periodicity of Metrical Structure., *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–499 (2021).

[10] Cheng, T., Fukayama, S. and Goto, M.: Joint beat and downbeat tracking based on CRNN models and a comparison of using different context ranges in convolutional layers, *the International Computer Music Conference (ICMC)* (2021).

[11] Hung, Y.-N., Wang, J.-C., Song, X., Lu, W.-T. and Won, M.: Modeling beats and downbeats with a time-frequency Transformer, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 401–405 (2022).

[12] Böck, S., Krebs, F. and Widmer, G.: A Multi-model Approach to Beat Tracking Considering Heterogeneous Music Styles, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 603–608 (2014).

[13] Krebs, F., Böck, S. and Widmer, G.: An Efficient State-Space Model for Joint Tempo and Meter Tracking, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 72–78 (2015).

[14] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F. and Widmer, G.: madmom: a new Python Audio and Music Signal Processing Library, *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 1174–1178 (2016).

[15] Davies, M. E. and Plumbley, M. D.: Context-dependent beat tracking of musical audio, *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 1009–1020 (2007).

[16] Ellis, D. P.: Beat tracking by dynamic programming, *Journal of New Music Research*, pp. 51–60 (2007).

[17] Klapuri, A. P., Eronen, A. J. and Astola, J. T.: Analysis of the meter of acoustic musical signals, *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 342–355 (2005).

[18] Peeters, G. and Papadopoulos, H.: Simultaneous beat and downbeat-tracking using a probabilistic framework: Theory and large-scale evaluation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 1754–1769 (2010).

[19] Böck, S. and Schedl, M.: Enhanced Beat Tracking with Context-Aware Neural Network, *International Conference on Digital Audio Effects (DAFx)*, pp. 135–139 (2011).

[20] Krebs, F., Böck, S., Dorfer, M. and Widmer, G.: Downbeat Tracking Using Beat Synchronous Features with Recurrent Neural Networks., *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 129–135 (2016).

[21] Bai, S., Kolter, J. Z. and Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, *arXiv preprint arXiv:1803.01271* (2018).

[22] Fuentes, M., McFee, B., Crayencour, H., Essid, S. and Bello, J.: Analysis of common design choices in deep learning systems for downbeat tracking, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 106–112.

[23] Fuentes, M., McFee, B., Crayencour, H. C., Essid, S. and Bello, J. P.: A music structure informed downbeat tracking system using skip-chain conditional random fields and deep learning, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 481–485 (2019).

[24] Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C. and Cano, P.: An experimental comparison of audio tempo induction algorithms, *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, pp. 1832–1844 (2006).

[25] Krebs, F., Böck, S. and Widmer, G.: Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 227–232 (2013).

[26] Hainsworth, S. W. and Macleod, M. D.: Particle Filtering Applied to Musical Tempo Tracking, *Journal on Advances in Signal Processing (EURASIP)*, pp. 2385–2395 (2004).

[27] Davies, M. E., Degara, N. and Plumbley, M. D.: Evaluation methods for musical audio beat tracking algorithms, *Queen Mary University of London, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06* (2009).

[28] Hockman, J., Davies, M. E. P. and Fujinaga, I.: One in the Jungle: Downbeat Detection in Hardcore, Jungle, and Drum and Bass, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 169–174 (2012).