1P-08

拍節構造の周期性に基づく深層ビート推定

大山 偉永[1]　石塚 崚斗[2]　錦見 亮[2]　吉井 和佳[2]

[1]京都大学 工学部情報学科　[2]京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

音楽音響信号に対する自動採譜や構造解析において，楽曲の拍節構造，すなわち拍子やビート・ダウンビート時刻を既知とした手法 [1,2] が提案されつつあり，拍節構造の自動推定は重要な課題となっている．特に，ビート推定に関しては多くの研究が存在し，現在では，深層ニューラルネットワーク (DNN) を用いて，音響特徴系列に対して，各時刻におけるビート存在確率を推定する方法が標準的となっている [5]．ただし，最終的にビート時刻を決定するには，推定されたビート存在確率系列からピークを検出する後処理が必要になる．しかし，DNN の学習時にビートの周期性（テンポに依存）を陽に考慮しない場合，ビート存在確率系列は明確な周期性を持たず，ピーク検出が困難になることがあった．

近年，ビート・ダウンビート・テンポの間には密接な関係があることから，これら要素のうち 2 つあるいは 3 つ全てを同時に推定する研究が行われている [3,6,9]．このアプローチでは，テンポ情報を考慮することにより，ビートやダウンビートの存在確率系列が周期性をもつよう誘導する効果が期待できる．しかし，通常のマルチタスク学習の枠組みでは，一つの DNN を途中で分岐させ，以降はタスクごとに独立した出力を行うため，タスク間の出力の一貫性が保証されない問題があった．

これらの問題を解決するため，本稿では，拍節構造の周期性を陽に考慮しつつ，一貫性を保った状態でビート・ダウンビート・テンポを同時に推定する手法を提案する．本手法の肝は，各時刻で，ビートおよびダウンビートの存在確率ではなく，それらの周期的な位相を出力するよう DNN を学習することにある．テンポは位相の時間変化（角速度）として微分可能なまま計算可能であり，全体を逆誤差伝播法で一挙に最適化可能になる．

2. 提案法

本章では，提案するビート・ダウンビート・テンポの同時推定手法を述べる（図 1）．F を周波数ビン数，T をフレーム数とする．入力はメルスペクトログラム $\mathbf{X} \in \mathbb{R}^{F\times T}$ であり，出力はビート時刻系列，ダウンビート時刻系列，テンポ (BPM) 系列 $\mathbf{Y} \triangleq \{y_t\}_{t=1}^{T}$ である．

中間表現として，ビート位相系列 $\mathbf{Z}^b \triangleq \{z_t^b\}_{t=1}^{T}$ と，ダウンビート位相系列 $\mathbf{Z}^d \triangleq \{z_t^d\}_{t=1}^{T}$ を定義する．ここで，$z_t^b, z_t^d \in [0, 2\pi)$ である．また，1 周期に相当する 2π を K クラスに量子化した系列をそれぞれ $\hat{\mathbf{Z}}^b \triangleq \{\hat{\mathbf{z}}_t^b\}_{t=1}^{T}$，$\hat{\mathbf{Z}}^d \triangleq \{\hat{\mathbf{z}}_t^d\}_{t=1}^{T}$ とする．ここで，$\hat{\mathbf{z}}_t^b \in \{0,1\}^K$ は，$\frac{2\pi(k-1)}{K} \le z_t^b < \frac{2\pi k}{K}$ である場合に，$\hat{z}_{tk}^b = 1$ となる one-hot ベクトルであり，$\hat{\mathbf{z}}_t^d$ も同様に定義する．k は k 番目

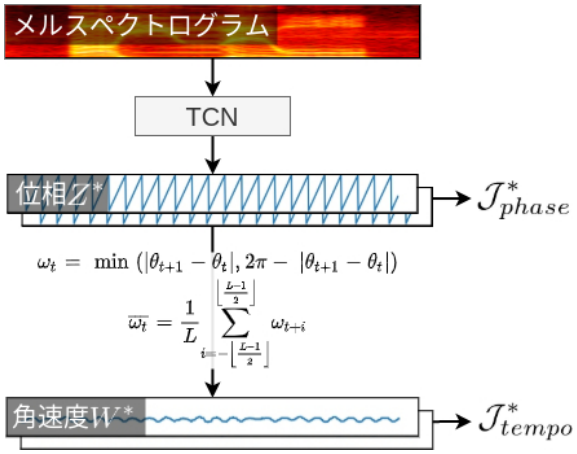

図 1: ビート・ダウンビート・テンポのマルチタスク学習

のクラスを表す．時間畳み込みネットワーク (TCN) [4,5] を用いて，$\mathbf{X}$ を入力として，各フレームのビートおよびダウンビートの位相を K クラスに分類する問題を解く．いま，TCN が出力する K クラスの確率系列をそれぞれ $\boldsymbol{\psi}^b \triangleq \{\boldsymbol{\psi}_t^b\}_{t=1}^{T}$ および $\boldsymbol{\psi}^d \triangleq \{\boldsymbol{\psi}_t^d\}_{t=1}^{T}$ とすると，正解データ $\hat{\mathbf{Z}}^b$ および $\hat{\mathbf{Z}}^b$ に対する対数尤度は次式で計算できる．

$$\mathcal{J}^*_{\text{phase}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \hat{z}_{tk}^* \log \psi_{tk}^* \tag{1}$$

ここで，* は ^{b} あるいは ^{d} である．

次に，TCN の出力である確率系列 $\boldsymbol{\psi}^b, \boldsymbol{\psi}^d$ を，微分可能なまま位相系列 $\mathbf{Z}^b, \mathbf{Z}^d$ にそれぞれ変換する．

$$z_t^* = \frac{2\pi}{K} \times \boldsymbol{a}^T \text{Gumbel-softmax}(\boldsymbol{\psi}_t^*) \tag{2}$$

ここで，$\boldsymbol{a} = [1, \ldots, K]^T$ はインデックスベクトルであり，Gumbel-softmax$(\boldsymbol{\psi}_t^*)$ は，離散分布 $\boldsymbol{\psi}_t^*$ に従う確率変数 $\hat{\boldsymbol{z}}_t^*$ をサンプリングする微分可能な関数である．

位相系列 $\mathbf{Z}^b, \mathbf{Z}^d$ を角速度系列 $\mathbf{W}^b \triangleq \{w_t^b\}_{t=1}^{T-1}$，$\mathbf{W}^d \triangleq \{w_t^d\}_{t=1}^{T-1}$ にそれぞれ変換する．ここで w_t^* は，位相の不連続点に注意して次式で定義される．

$$w_t^* = \min(|z_{t+1}^* - z_t^*|, 2\pi - |z_{t+1}^* - z_t^*|) \tag{3}$$

一般に，高周波成分（ノイズ）を含む系列に対する微分操作では，ノイズが強調される．そこで，ノイズの影響を軽減するため，瞬時的な角速度 w_t^* の代わりに，次式で定義される平均角速度系列 $\overline{w}_t^*$ を用いる．

$$\overline{w}_t^* = \frac{1}{L}\sum_{i=-\lfloor\frac{L-1}{2}\rfloor}^{\lfloor\frac{L-1}{2}\rfloor} w_{t+i}^* \tag{4}$$

ここで，$\lfloor\cdot\rfloor$ は床関数，$L \in \mathbb{Z}^+$ はカーネルサイズを表す．

テンポ y_t は 2 通りの計算が可能で，y_t^b, y_t^d とする．

$$y_t^b = \frac{60\times\text{fps}}{2\pi}\overline{w}_t^b,\ y_t^d = \frac{60\times\text{fps}}{2\pi}\omega_t^d \times \text{bpb} \tag{5}$$

Deep Beat Estimation Based on Periodicity of Metrical Structure: *Takehisa Ooyama, Ryoto Ishizuka, Ryo Nishikimi, and Kazuyoshi Yoshii (Kyoto Univ.)*

表 1: 評価結果 (%)． b, d, t はそれぞれビート，ダウンビート，テンポを表し，従来法 (bdt) は従来法を用いてビート，ダウンビート，テンポの同時推定を行ったことを表す．

	ビート					ダウンビート					テンポ	
手法	F 値	CMLc	CMLt	AMLc	AMLt	F 値	CMLc	CMLt	AMLc	AMLt	acc1	acc2
従来法 (bt)	**92.7**	70.9	79.6	78.6	**90.7**	-	-	-	-	-	69.7	81.8
提案法 (bt)	91.1	76.2	81.5	**81.5**	90.3	-	-	-	-	-	90.9	**98.5**
従来法 (dt)	-	-	-	-	-	85.6	76.8	79.8	79.6	84.4	84.8	90.9
提案法 (dt)	-	-	-	-	-	83.6	76.7	79.2	**82.1**	**86.5**	**98.5**	**98.5**
従来法 (bdt)	91.7	65.4	77.7	71.7	89.4	78.4	60.7	69.8	63.5	73.8	66.7	74.2
提案法 (bdt)	92.4	**77.7**	**83.1**	80.8	89.3	**85.8**	**76.9**	**80.2**	81.5	86.1	95.5	**98.5**

ここで，fps は 1 秒あたりのフレーム数 (frames per second)，bpb は 1 ダウンビートあたりのビート数 (beats per bar) を表す．

式 (5) に基づいて，正解テンポ $\hat{\mathbf{Y}}$ から計算した角速度の正解ラベル $\hat{\mathbf{W}}^b \triangleq \{\hat{w}_t^b\}_{t=1}^{T-1}$，$\hat{\mathbf{W}}^d \triangleq \{\hat{w}_t^d\}_{t=1}^{T-1}$ に対してガウス分布に基づく対数尤度を計算する．

$$\mathcal{J}^*_{\mathrm{tempo}} = -\sqrt{\frac{\sum_{t=1}^{T-1}(w_t^* - \hat{w}_t^*)^2}{T-1}} \tag{6}$$

式 (5) より，$\mathcal{J}^*_{\mathrm{tempo}}$ の最大化はテンポに関するガウス尤度最大化と等価である．TCN のパラメーターは以下の重み付け和を最大化するように学習する．

$$\mathcal{J}_{\mathrm{total}} = \mathcal{J}^b_{\mathrm{phase}} + \mathcal{J}^d_{\mathrm{phase}} + \alpha(\mathcal{J}^b_{\mathrm{tempo}} + \mathcal{J}^d_{\mathrm{tempo}}) \tag{7}$$

ここで，$\alpha \in \mathbb{R}+$ はテンポに関する重みである．$\mathcal{J}^*_{\mathrm{phase}}$ と $\mathcal{J}^*_{\mathrm{tempo}}$ を同時に最大化することで，位相の変化率が一定になり，テンポと位相の整合性が向上することが期待される．

最終的なビート時刻とダウンビート時刻系列は，単純なピーク検出を行うことで得られる．従来法で用いられている Dynamic Bayesian Network (DBN) [7] に基づくピーク検出は，連続的に変換する位相を出力する本提案手法には適さないため，今回は使用しなかった．曲全体のテンポは，TCN の出力する $\mathbf{Z}^b$ に対してフーリエ変換を施し，式 (5) に基づいて推定する．

3. 評価実験

実験には，RWC ポピュラー音楽データベース [8] に含まれる 4/4 拍子の 87 曲を使用した．学習データには 87 曲からランダムに選んだ 70 曲を使用し，残りの 17 曲をテストデータとした．学習データを拡張するため，各曲に対して-12 半音から+12 半音の間で半音ごとにピッチシフトを行い，$[0.5, 2]$ からランダムに選んだ 5 つの倍率に基づいて時間伸縮を行った．

ビート・テンポの推定，ダウンビート・テンポの推定，ビート・ダウンビート・テンポの推定の 3 つの実験を提案手法と従来法 [9] に対してそれぞれ行い比較した．従来法 [9] は，ビート・ダウンビート・テンポの同時推定は行えないので，[3] に基づき，3 つの推定を同時に行えるように変更を加えた．

ビートの評価尺度として [10] に定義されている F 値，推定したビート間隔の一定度合いを表す CMLc と CMLt，それに加え 2 倍や 1/2 倍などの複数の時間間隔も正解とする AMLc，AMLt を用いた．テンポの評価尺度として [9] に従い，$\pm 4\%$ の誤差を許容する acc1 と，それに加え 2 倍や 1/2 倍などのテンポも許容する acc2 を用いた．

表 1 に評価結果を示す．評価指標 CMLc，CMLt，AMLc，AMLt において，提案手法が平均的に従来法よりも高い精度を示した．角速度誤差を最小化することで位相の変化率が滑らかになり，一定の間隔でのビート検出が可能になったと考えられる．また提案手法では，ビート・ダウンビート・テンポを同時に学習することで，これらのうち 2 つで学習したモデルよりも平均的に高い精度を示した．

4. おわりに

本稿では，ビートとダウンビートの周期性を考慮したビート・ダウンビート・テンポの同時推定手法を提案した．今後は，角速度系列を計算したあとで移動平均を計算する代わりに，あらかじめ平滑化フィルタ（Savitzky-Golay 法など）を位相系列に適用したのち，角速度系列を計算するアプローチとの比較を行う．また，位相をフーリエ変換することで計算されるテンポに対し，逆フーリエ変換を適用することでノイズを除去した位相を推定する方法についても検討する予定である．

謝辞 本研究の一部は，JST ACCEL No. JPM-JAC1602，JSPS 科研費 No. 16H01744, No. 19H04137 の支援を受けた．

参考文献

[1] G. Shibata *et al.*: “Music Structure Analysis Based on an LSTM-HSMM Hybrid Model,” *ISMIR*, 2020.

[2] R. Nishikimi *et al.*: “Bayesian Singing Transcription Based on a Hierarchical Generative Model of Keys, Musical Notes, and F0 Trajectories,” *TASLP*, 2020.

[3] S. Böck *et al.*: “Deconstruct, Analyse, Reconstruct: How To Improve Tempo, Beat, and Downbeat Estimation,” *ISMIR*, 2020.

[4] S. Bai *et al.*: “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” arXiv:1803.01271v2, 2018.

[5] S. Böck *et al.*: “Temporal Convolutional Networks for Musical Audio Beat Tracking,” *EUSIPCO*, 2019.

[6] S. Böck *et al.*: “Joint Beat and Downbeat Tracking with Recurrent Neural Networks,” *ISMIR*, 2016.

[7] F. Krebs *et al.*: “An Efficient State-Space Model for Joint Tempo and Meter Tracking,” *ISMIR*, 2015.

[8] M. Goto *et al.*: “RWC Music Database: Popular, Classical and Jazz Music Databases,” *ISMIR*, 2002.

[9] S. Böck *et al.*: “Multi-Task Learning of Tempo and Beat: Learning One to Improve the Other,” *ISMIR*, 2019.

[10] M. Davies *et al.*: “Evaluation Methods for Musical Audio Beat Tracking Algorithms,” Queen Mary University of London, Centre for Digital Music, Technical Report C4DM-TR-09-06, 2009.