4T-07

マルチタスク学習に基づくドラム採譜と拍節構造推定

鎌倉 大地[1]　大山 偉永[2]　吉井 和佳[2,3]

[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　[3]JST さきがけ

1. はじめに

ドラム採譜は，音響音楽信号からドラムパート（本稿ではバスドラム (BD)・スネアドラム (SD)・ハイハット (HH) について扱う）の楽譜を推定するタスクである．最近の標準的なアプローチでは，深層ニューラルネットワーク（DNN）を用いてドラムの発音時刻を検出したのち，別途推定した拍節構造（ビート・ダウンビート時刻）を用いて量子化を行うことで，楽譜上の位置を特定する．ここで，ドラムはビート・ダウンビート上で発音しやすく，ドラムパターンはダウンビート単位で反復される傾向があることから，ドラム採譜と拍節構造推定のマルチタスク (MT) 学習が有効である [1]．しかし，音響信号からタスク共通の潜在特徴を抽出したのち，タスク固有のネットワークに入力する一般的な分枝型アーキテクチャでは，ドラム・ビート・ダウンビートそれぞれが持つ周期性や，最終的な出力における要素間の同期性を十分に考慮できているわけではなかった．

この問題を解決するため，本研究では，ドラム・ビート・ダウンビート時刻系列それぞれの自己相関，および各組の相互相関を高めるような正則化を導入したマルチタスク学習法を提案する．具体的には，自己（相互）相関関数の離散フーリエ変換 (DFT) であるオート（クロス）スペクトルにおいて，一部の周波数にエネルギーが集中しているほど，すなわちエントロピーが小さいほど周期性（同期性）が高いことを利用する．分枝型 DNN の学習時には，各要素の教師データに対する誤差とこれらエントロピーの和を最小化する．

2. 提案法

本章では，提案するドラム採譜と拍節構造推定のマルチタスク学習法について説明する（図 1）．

2.1 問題設定

音響信号の左右チャネルのパワースペクトログラム $\mathbf{X} \in \mathbb{R}^{2\times F\times T}$ に対し，各フレームにおけるドラム・ビート・ダウンビートの有無を表す $\mathbf{Y}^{\mathrm{D}} \in \{0,1\}^{K\times T}$，$\mathbf{Y}^{\mathrm{B}} \in \{0,1\}^{T}$，$\mathbf{Y}^{\mathrm{W}} \in \{0,1\}^{T}$ を推定する．ここで，F は周波数ビン数，T はフレーム数，K はドラムの種類数である（本稿では $K=3$）．ドラム採譜では本来，拍節構造に基づく量子化が必要であるが，本稿では紙面の都合上，フレーム単位での推定・評価に絞って報告する．

2.2 教師ありマルチタスク学習

本研究では，潜在特徴を抽出する共通の畳み込みニューラルネットワーク (CNN) に対し，ドラム検出には双方向長・短期記憶ネットワーク (BLSTM)，ビート・ダウンビート検出には時間畳み込みネットワーク (TCN) [3,4] を接続した分枝型 DNN を用いた．本 DNN は，$\mathbf{X}$ を入力として，各フレームにおけるドラム・ビート・ダウンビートの存在確率 $\boldsymbol{\phi}^{\mathrm{D}} \in [0,1]^{K\times T}$，$\boldsymbol{\phi}^{\mathrm{B}} \in [0,1]^{T}$，$\boldsymbol{\phi}^{\mathrm{W}} \in [0,1]^{T}$ を出力する．このとき，各要素の教師データ $\hat{\mathbf{Y}}^{\mathrm{D}} \in \{0,1\}^{K\times T}$，$\hat{\mathbf{Y}}^{\mathrm{B}} \in \{0,1\}^{T}$，$\hat{\mathbf{Y}}^{\mathrm{W}} \in \{0,1\}^{T}$ に対してバイナリクロスエントロピー (BCE) に基づく誤差関数が計算できる．

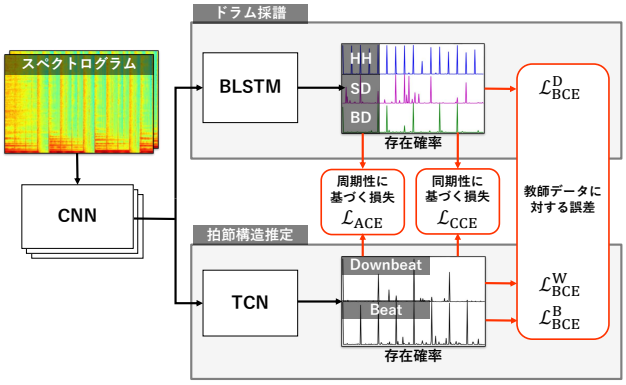

図 1: ドラム採譜と拍節構造推定のマルチタスク学習

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{BCE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T}\left(\hat{Y}^{\mathrm{D}}_{kt}\log\phi^{\mathrm{D}}_{kt} + (1-\hat{Y}^{\mathrm{D}}_{kt})\log(1-\phi^{\mathrm{D}}_{kt})\right) \quad (1)$$

$\mathcal{L}^{\mathrm{B}}_{\mathrm{BCE}}$ および $\mathcal{L}^{\mathrm{W}}_{\mathrm{BCE}}$ も同様に定める．マルチタスク学習では，これらの重み付き和を最小化することが基本となる．

2.3 周期性・同期性に基づく正規化

標準的な教師あり学習に対し，ドラム・ビート・ダウンビートの周期性・同期性に基づく正規化を導入する．まず，ドラムの存在確率系列 $\boldsymbol{\phi}^{\mathrm{D}}$ の自己相関関数のスペクトルのエントロピー (ACE) に基づく損失関数を考える．

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{ACE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T}\tilde{\phi}^{\mathrm{D}}_{kt}\log\tilde{\phi}^{\mathrm{D}}_{kt} \quad (2)$$

$$\tilde{\boldsymbol{\phi}}^{\mathrm{D}}_{k} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{D}}_{k}|^{\cdot 2}) \quad (3)$$

ここで，$F \in \mathbb{C}^{T\times T}$ は DFT 行列，$\boldsymbol{\phi}^{\mathrm{D}}_{k} \in [0,1]^{T}$ はドラム k の存在確率系列，ベクトル $\mathbf{x}$ に対し，$|\mathbf{x}|$ は各要素の絶対値，$\mathbf{x}^{\cdot 2}$ は各要素の二乗，$\mathrm{Normalize}(\mathbf{x})$ は L1 ノルムが 1 となる正規化を表す．式 (3) では，ウィナー・ヒンチンの定理により，$\boldsymbol{\phi}^{\mathrm{D}}$ の自己相関関数のスペクトルが $\boldsymbol{\phi}^{\mathrm{D}}$ のパワースペクトルと等価であることを用いた．ビート・ダウンビートの存在確率系列 $\boldsymbol{\phi}^{\mathrm{B}}$，$\boldsymbol{\phi}^{\mathrm{W}}$ に関する損失関数 $\mathcal{L}^{\mathrm{B}}_{\mathrm{ACE}}$，$\mathcal{L}^{\mathrm{W}}_{\mathrm{ACE}}$ も同様に定める．

次に，ドラムの存在確率系列 $\boldsymbol{\phi}^{\mathrm{D}}$ およびビートの存在確率系列 $\boldsymbol{\phi}^{\mathrm{B}}$ の相互相関関数のスペクトルのエントロピー (CCE) に基づく損失関数を考える．

$$\mathcal{L}^{\mathrm{DB}}_{\mathrm{CCE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T}\tilde{\phi}^{\mathrm{DB}}_{kt}\log\tilde{\phi}^{\mathrm{DB}}_{kt} \quad (4)$$

$$\tilde{\boldsymbol{\phi}}^{\mathrm{DB}}_{k} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{D}}_{k}||F\boldsymbol{\phi}^{\mathrm{B}}|) \quad (5)$$

式 (5) についても，ウィナー・ヒンチンの定理により，二系列の相互相関関数のスペクトルが両者のクロススペク

Drum Transcription and Metrical Structure Estimation Based on Multitask Learning: *Daichi Kamakura, Takehisa Oyama, and Kazuyoshi Yoshii (Kyoto Univ.)*

表 1: RWC ポピュラー音楽データベース中 64 曲に対する検出精度（F 値）．D はドラム，B はビート，W はダウンビートを表す．太字は最良値から 0.5 pts 以内の値を示す．

MT 学習			正則化		ドラム				拍節構造	
D	B	W	ACE	CCE	BD	SD	HH	Ave.	B	W
	✓	✓			-	-	-	-	90.5	87.3
✓					**78.0**	71.2	74.2	74.5	-	-
✓			✓		77.0	70.1	74.7	73.9	-	-
✓	✓				76.6	69.6	76.5	74.2	90.6	-
✓	✓		✓		76.6	71.2	75.2	74.3	90.3	-
✓	✓			✓	77.0	71.3	75.1	74.5	88.9	-
✓	✓		✓	✓	**77.6**	**73.5**	75.7	**75.6**	89.5	-
✓	✓	✓			75.1	68.7	72.4	72.1	93.9	**89.2**
✓	✓	✓	✓		77.1	74.1	73.7	75.0	**94.6**	**89.1**
✓	✓	✓		✓	74.2	70.7	76.4	73.8	**94.1**	87.2
✓	✓	✓	✓	✓	76.8	72.2	**77.1**	**75.4**	92.1	86.3

トルと等価であることを用いた．ドラム・ダウンビートの CCE に基づく損失関数 $\mathcal{L}_{\mathrm{CCE}}^{\mathrm{DW}}$ およびビート・ダウンビートの CCE に基づく損失関数 $\mathcal{L}_{\mathrm{CCE}}^{\mathrm{BW}}$ も同様に定める．

最終的に，最小化すべき損失関数 $\mathcal{L}$ は次式で定まる．

$$\mathcal{L} = \lambda_{\mathrm{BCE}}\mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{ACE}}\mathcal{L}_{\mathrm{ACE}} + \lambda_{\mathrm{CCE}}\mathcal{L}_{\mathrm{CCE}} \quad (6)$$

ここで，$\mathcal{L}_{\mathrm{BCE}}$，$\mathcal{L}_{\mathrm{ACE}}$ および $\mathcal{L}_{\mathrm{CCE}}$ は以下で与えられる．

$$\mathcal{L}_{\mathrm{BCE}} = \lambda_{\mathrm{BCE}}^{\mathrm{D}}\mathcal{L}_{\mathrm{BCE}}^{\mathrm{D}} + \lambda_{\mathrm{BCE}}^{\mathrm{B}}\mathcal{L}_{\mathrm{BCE}}^{\mathrm{B}} + \lambda_{\mathrm{BCE}}^{\mathrm{W}}\mathcal{L}_{\mathrm{BCE}}^{\mathrm{W}} \quad (7)$$

$$\mathcal{L}_{\mathrm{ACE}} = \lambda_{\mathrm{ACE}}^{\mathrm{D}}\mathcal{L}_{\mathrm{ACE}}^{\mathrm{D}} + \lambda_{\mathrm{ACE}}^{\mathrm{B}}\mathcal{L}_{\mathrm{ACE}}^{\mathrm{B}} + \lambda_{\mathrm{ACE}}^{\mathrm{W}}\mathcal{L}_{\mathrm{ACE}}^{\mathrm{W}} \quad (8)$$

$$\mathcal{L}_{\mathrm{CCE}} = \lambda_{\mathrm{CCE}}^{\mathrm{DB}}\mathcal{L}_{\mathrm{CCE}}^{\mathrm{DB}} + \lambda_{\mathrm{CCE}}^{\mathrm{DW}}\mathcal{L}_{\mathrm{CCE}}^{\mathrm{DW}} + \lambda_{\mathrm{CCE}}^{\mathrm{BW}}\mathcal{L}_{\mathrm{CCE}}^{\mathrm{BW}} \quad (9)$$

ここで，$\lambda_{\mathrm{BCE,ACE}}^{\{\mathrm{D,\,B,\,W}\}}$ および $\lambda_{\mathrm{CCE}}^{\{\mathrm{DB,\,DW,\,BW}\}}$ は重みである．

2.4 存在確率系列に対するピーク検出

最終的な出力 $\mathbf{Y}^{\mathrm{D}}$，$\mathbf{Y}^{\mathrm{B}}$，$\mathbf{Y}^{\mathrm{W}}$ は ϕ^{D}，ϕ^{B}，ϕ^{W} を二値分類することで得られるが，単純な閾値処理では性能に限界がある．そこで，ビート時刻系列とダウンビート時刻系列は，周期性を考慮した Dynamic Bayesian Network (DBN) [5] を用いたピーク検出を行う．

3. 評価実験

実験には，RWC ポピュラー音楽データベース [6] のうち，ドラムを含む 64 曲（発音時刻のアノテーションが正確なもの）と含まない 10 曲を使用した．サンプリング周波数 44.1 kHz のステレオ信号に窓幅 1024 点・シフト幅 441 点の短時間フーリエ変換 (STFT) を適用し，左右チャネルを連結して $\mathbf{X}$ とした．正解の発音時刻は，量子化に伴う丸め誤差や，アノテーションのゆらぎの影響を受けている．そのため，学習時には平均 0 ms・標準偏差 12 ms の正規乱数に従う摂動を加えることで教師データとした．また，各楽曲に対し，平均 1・標準偏差 0.1 の正規乱数に従う倍率で時間伸縮を行った．提案法における損失関数の重みは，$\lambda_{\mathrm{BCE}}{=}\lambda_{\mathrm{ACE}}{=}\lambda_{\mathrm{CCE}}{=}1$，$\lambda_{\mathrm{BCE}}^{\mathrm{D}}{=}1$，$\lambda_{\mathrm{BCE}}^{\mathrm{B}}{=}\lambda_{\mathrm{BCE}}^{\mathrm{W}}{=}0.1$，$\lambda_{\mathrm{ACE}}^{\mathrm{D}}{=}1$，$\lambda_{\mathrm{ACE}}^{\mathrm{B}}{=}\lambda_{\mathrm{ACE}}^{\mathrm{W}}{=}0$，$\lambda_{\mathrm{CCE}}^{\mathrm{DB}}{=}\lambda_{\mathrm{CCE}}^{\mathrm{DW}}{=}\lambda_{\mathrm{CCE}}^{\mathrm{BW}}{=}1$ とした．

提案法におけるマルチタスク学習と周期性・同期性に基づく正則化の有効性を検証するため，ドラム・ビート・ダウンビート検出をそれぞれ独立で行うか，二要素のみをマルチタスク学習する場合と，ACE・CCE に基づく損失関数を用いない場合とで比較した．五分割交差検証を行い，ドラムを含む 64 曲に対して評価を行った．

表 2: RWC ポピュラー音楽データベース全 100 曲に対するビート・ダウンビート検出精度．

	ビート検出			ダウンビート検出		
	F 値	CMLt	AMLt	F 値	CMLt	AMLt
既存法 [2]	89.5	81.8	91.5	83.1	80.3	88.4
提案法	92.8	85.9	94.3	86.3	81.3	89.4

評価尺度には F 値を用い，許容誤差はドラム検出については 30 ms，ビート・ダウンビート検出については 70 ms とした．また，ビート間隔の安定性を表す CMLt に加えて，正解テンポ以外に倍・半テンポなどの解釈も考慮したうえでの安定性を示す AMLt を用いた [7]．

表 1 に結果を示す．三要素のマルチタスク学習により，ドラム検出とビート・ダウンビート検出を独立で行うより，ビート検出で約 4 pts，ダウンビート検出で約 2 pts 改善が見られた．周期性と同期性に基づく正則化を同時に用いると，ドラム検出で約 1〜3 pts の改善が見られた．一方，ビート検出では，ピーク検出における周期性を考慮した DBN の効果が大きく，改善が見られなかった．ただし，単純な閾値処理を用いる場合には改善が見られた．

表 1 でビートの検出精度が最も高かった設定（下から 3 行目に対応）で，RWC ポピュラー音楽データベース中の全 100 曲に対する結果を表 2 に示す．既存法 [2] に基づく推定と比べて約 3 pts 高い精度を示した．

4. おわりに

本稿では，ドラム・ビート・ダウンビート検出において，周期性と同期性に基づく正則化を導入したマルチタスク学習法を提案した．実験から，マルチタスク学習と各正則化の有効性が確認できたが，各損失関数の適切な重みの設定や，提案法の性能を最大限引き出すためのピーク検出方法についてはさらなる研究が必要である．

今後，楽曲中でのテンポや拍子の変化（非定常性）に対応するため，各要素の存在確率系列に対して DFT ではなく STFT を適用し，楽曲全体ではなくフレーム単位での周期性・同期性に基づく正則化を検証する．この際，テンポを考慮した窓幅やシフト幅の設定について検討を行う．また，拍節構造の周期性を考慮する別の方法として，存在確率系列のかわりに正弦波や三角波などの周期関数を出力する方式 [8] についても検討する．

謝辞 本研究の一部は，JST PRESTO No. JPMJPR20CB および科研費 No. 19H04137, 21H03572 の支援を受けた．

参考文献

[1] R. Vogl *et al.*: “Drum Transcription via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks,” *ISMIR*, 2017.
[2] S. Böck *et al.*: “Deconstruct, Analyse, Reconstruct: How to Improve Tempo, Beat, and Downbeat Estimation,” *ISMIR*, 2020.
[3] S. Bai *et al.*: “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” arXiv:1803.01271v2, 2018.
[4] S. Böck *et al.*: “Temporal Convolutional Networks for Musical Audio Beat Tracking,” *EUSIPCO*, 2019.
[5] F. Krebs *et al.*: “An Efficient State-Space Model for Joint Tempo and Meter Tracking,” *ISMIR*, 2015.
[6] M. Goto *et al.*: “RWC Music Database: Popular, Classical and Jazz Music Databases,” *ISMIR*, 2002.
[7] 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.
[8] T. Oyama *et al.*: “Phase-Aware Joint Beat and Downbeat Estimation Based on Periodicity of Metrical Structure,” *ISMIR*, 2021.