Vol.2021-MUS-131 No.65
Vol.2021-SLP-137 No.65
2021/6/19

複数パート間のズレを含む演奏音に対するマルチパートビートトラッキング

福谷 和貴[1]　酒向 慎司[1,a)]

概要：本研究では、楽器ごとに一つのビートラベル列が存在し、複数のビートラベル列を持つ音楽演奏を対象とし、その混合音に対して複数パートのビート位置を同時にトラッキングすることを試み、このようなマルチパートビートトラッキングのための新たな手法を提案した。音源分離手法によって分離した単独音に対するビートトラッキングを組み合わせた方法と比較することで提案手法の有効性を確認した。

1. はじめに

合奏の際，一定のリズムで演奏するには相応の技量が必要であり，初心者にとっては困難なことである．そこで，合奏において楽器ごとの演奏タイミングを取得し，適切な方法で演奏者に示すことで練習支援ができれば有益であると考えられる．演奏タイミングを演奏者に示すことに着目した演奏支援システムに関する研究は従来行われているが，練習支援の対象として電子楽器が用いられることが多く，その演奏情報と楽譜とを比較し，その結果を視覚的に演奏者に示すことで練習支援を実現している．電子楽器を用いることで特定の楽器の演奏タイミングを簡便に，また正確に取得することができるが，楽器の種類が限定されるため，対象楽器が限られる点が問題である．また，この時，合奏において個々のパートの演奏タイミングがずれており，異なる複数の演奏タイミングを示すことに関して有効な手法が必要である．

そこで本研究では音響信号からのビートトラッキングを利用した練習支援に着目する．合奏において楽器ごとのビートラベル列を認識することで，楽器ごとの演奏タイミングを示すことを目標とする．しかし，ビートトラッキングの一般的な問題設定は，一つの演奏に対してビートラベル列が一つである．楽器ごとに一つのビートラベル列が存在し，複数のビートラベル列を持つ音楽に対して行うビートトラッキングのアプローチとして，音源分離を行い，楽器ごとにビートトラッキングを行う方法と，混合音に対して複数パートのビート位置を同時にトラッキングする方法が考えられる．本研究では，後者のようなアプローチをマルチパートビートトラッキングと呼称し，このアプローチによる手法を提案する．筆者の知る限り音響信号から楽器ごとのビートラベル列を得る様なマルチパートのビートトラッキングは存在しない．従って，研究の第一歩として，本研究の問題設定では対象をギターとドラムの 2 つの楽器での演奏に限定する．また，ギターの演奏のみが演奏タイミングのズレを含む演奏であるとする．

[1] 名古屋工業大学
Nagoya Institute of Technology, Gokisocho, Showa-ku, Nagoya, 466–8555, Japan
[a)] s.sako@nitech.ac.jp

2. ビートトラッキング

ビートトラッキングはこれまでにも盛んに研究が進められており，音楽情報処理技術に関する国際コンペティションである MIREX (the Music Information Retrieval Evaluation eXchange) の主要なタスク (Audio Beat Tracking) の一つである [1]．一般的に，ビートは周期性を持っているものであり，楽曲のテンポを推定するテンポ推定とビートトラッキングは密接な関係にある．ビートトラッキングの基本的な手法では，音楽音響信号から，スペクトルなどの特徴量を抽出する．次に，抽出した特徴量から周期性を推定することで楽曲のテンポを推定する．テンポ推定では，この推定されたテンポを出力する．ビートトラッキングにおいては，抽出した特徴量のピーク位置と推定したテンポを考慮しビート位置が推定される．

2.1 ビートトラッキングに関する研究事例

信号処理による特徴量抽出やビート位置の周期性などを考慮したアルゴリズムがあり，その中の一例として後藤らの研究がある [2]．このシステムでは，周波数解析によりオンセット時刻を検出する．そのオンセット時刻を用いて，それぞれが異なるビート推定方法を持つ複数のエージェン

ト (識別機) がビート位置を推定する．その中から，最も信頼性の高いものを出力として選択することでビート位置を推定している．最近では，ニューラルネットワークを用いた手法が主流になってきている．Elowsson は音楽の周期性や，テンポなどをモデル化した複数のフィードフォワードニューラルネットワークを用いることで，ビート位置の推定を行っている [3].

上記のようなシンプルなニューラルネットワーク以外にも，時系列を考慮できる Recurrent Neural Network (RNN) や，RNN の一種であり，長期的な依存関係を学習することのできる Long short-term memory (LSTM) を用いた研究が多くなされている．Böck らは 3 層の Bidirectional LSTM (BLSTM) を用いたビートトラッキング手法をいくつか提案している [4], [5], [6]．Böck らの研究では，メルスペクトログラムを入力特徴量として BLSTM に入力し，楽曲のフレームごとのビートである確率を示す beat activation 関数を算出する．この関数と自己相関関数や Dynamic Bayesian Network を使用することでビート位置を推定している．この手法では，[6] では Ballroom dataset[7] において，F 値で 0.938 という精度が示されている．このデータセットは社交ダンスで使われる曲を収録しているデータセットであり，安定したテンポと，比較的分かりやすいビート位置が特徴である．一方，曲中で拍子記号が変わる，テンポが変動するなど，ビートトラッキングが困難な曲が収録された SMC dataset[8] に対しては，F 値で 0.516 という精度が示されている．

そのほか，Fiocchi らは Böck らの提案しているモデル [6] を用いて転移学習を行うことでギリシャ民謡へのビートトラッキングシステムを提案している [9].

2.2 本研究のねらい

一般的なビートトラッキングでは複数の楽器は同期して演奏され，1 曲につきビートラベル列は 1 つのみ存在する．しかし，本研究では演奏タイミングが同期されておらず，楽器ごとにビートラベル列が存在している曲に対して，マルチパートビートトラッキングを行う．マルチパートに対応するためには，音源数や編成も含めて同定する必要が出てくるが，本研究では音源数や編成などは既知とし，ギターとドラムの 2 種類の演奏に限定しているため，1 曲にギターとドラムの 2 つのビートラベル列が存在する状況を想定している．

このようなビートトラッキングのアプローチとして，音を楽器ごとに分離し，楽器ごとにビートトラッキングを行う方法と，混合音のままマルチパートビートトラッキングを行う方法が考えられる．しかし音源分離の性能に依存するほか，分離する楽器が限定されているなどの制約もある．従って本研究では，従来手法をマルチパートに拡張することを考える．MIREX で好成績を収めていることから，本研究においては，LSTM を用いた手法を採用しマルチパートに拡張する．本研究で対象としているギター，ドラムを対象に含んでいる音源分離手法として，Harmonic-percussive source separation (HPSS)[10] の他，近年の DNN に基づいたものが数多く提案されている．本研究では U-Net を利用した最新のモデルの一つとして Spleeter[11] を採用し，音源分離の手法と通常のビートトラッキング手法との組み合わせとマルチパートビートトラッキング手法を比較することで，マルチパート化の有効性を確認する．

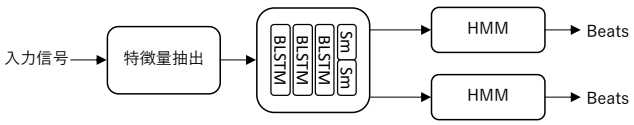

図 1 提案手法の概要図

3. 提案手法

本研究では BLSTM を用いた先行研究 [6] を参考にマルチパートビートトラッキング手法を提案する．1 に提案手法の概要図を示す．提案手法の入力はシングルチャンネルの音響信号である．先行研究は，出力は入力音響信号に対応する単一系列つの出力だが，本研究では，出力は入力音響信号に対応するギターパートとドラムパートのビートラベル列である．なお，提案モデルは簡単のため最小構成である 2 パートで実験を行っているが，原理的には複数のパートでのマルチパートビートトラッキングが可能な枠組みである．

3.1 特徴量抽出

BLSTM の入力として異なる窓幅で算出したメルスペクトログラムを連結したものを用いる．これは先行研究でも効果が確認されているものである．特徴量を算出する際の窓幅を 4,096，2,048，1,024 サンプル (92.8，46.4，23.2 ms) とし，シフト幅を 441 サンプル (10 ms) とする．3 種類の窓幅で算出される特徴量において，1 オクターブあたり 12，6，3 のバンド数を使用する．また，動的特徴量としてメルスペクトログラムの 1 次差を特徴量に連結する．結果として 381 次元の入力特徴量が算出される．

3.2 BLSTM の詳細

本研究ではニューラルネットワークのアーキテクチャとして LSTM を時間軸方向へ双方向に発展させた BLSTM を採用する．BLSTM は入力データの時間的前後関係をモデル化できるため，ビートトラッキング問題に適用可能である [4]．まず，マルチパート拡張する前の BLSTM について説明する．本研究では 3 層の BLSTM を用い，隠れ層は 1 層につき 25 の LSTM ユニットを持つ．入力層が 381 次元であり，出力層は beat と non-beat という 2 種類のラ

ベル情報に相当する 2 次元の層である．活性化関数として Softmax を用い，BLSTM の出力はあるフレームにおける，そのフレームがビートである確率である．

BLSTM の出力を時系列に並べたもの (beat activation 関数) と [12] で提案されている隠れマルコフモデル (hidden Markov model; HMM) を用いることでビート位置を推定する．この HMM では，隠れ状態はあるフレームにおけるテンポと小節内の拍の位置の 2 つの隠れ変数である．beat activation 関数が観測系列であり，Viterbi アルゴリズムによって最適な状態遷移系列を推定することで，ビート位置が推定される．本研究では madmom[13] を使って HMM を実装する．

3.3 BLSTM のマルチパートへの拡張

本研究では問題の簡単化のため，ギター演奏にズレが含まれており，ドラム演奏にはズレを含まない演奏が入力されることを想定している．一般的に，ドラム演奏が曲のリズムを担っており，ギター演奏だけでは，ビートトラッキングを期待通りに行えないことが予想される．また，ギター演奏には演奏タイミングのズレが存在することから，曲全体のビートの周期から外れたビートが存在するため，曲全体のビートの周期性を学習することが困難であると考えられる。

そこで，ドラムとギターを同時に学習に用いる方法を考える．基準となるドラムのビート位置を一緒に学習することで，曲全体のビートの周期性を学習することができ，ズレを含んだギター演奏に対しては，基準からのズレ具合のみを学習するだけで良いため，ギター演奏に対してビート位置を推定することが期待される [14].

図 2 は入力特徴系列のある 1 フレームにおける，ギターパート，ドラムパートのビート情報を推定するネットワークの概要図である．381 次元の入力層，各 25 ユニットの BLSTM 層が 3 層ある．この図のように BLSTM の出力層をギターとドラムのそれぞれの個別の出力層を用いることでマルチパートに拡張する．そして，ギターとドラムについての出力をそれぞれ時系列に並べ，それぞれを HMM に入力することで，ギター，ドラムそれぞれのビートラベル列を得る．

3.4 BLSTM の学習

学習は教師付き学習と早期停止を用いて行う．データセットに対して 5 倍の交差検証を行う．学習データの 2 割を検証データとして使用し，残りのデータで学習し，バッチサイズを 8 としたミニバッチ学習を行う．ネットワークの隠れ状態を [−0.1, 0.1] の範囲でランダムに初期化する．最適化関数として Adam を用い，学習率を 0.01 とする．損失関数としてクロスエントロピーを用いる．過学習を防ぐために，20 エポックの間に検証データに対する損失の改善が見られない場合，学習を終了する．その後，学習率を 0.001 に変更し，同じ終了条件において再学習する．

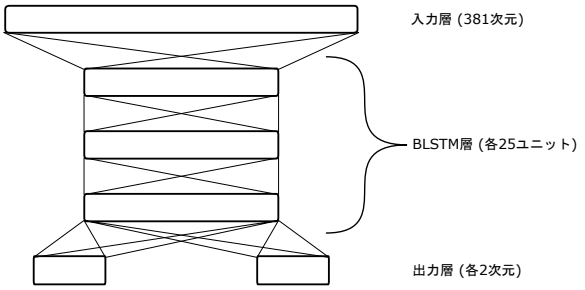

図 2 提案モデルの概要図

4. データセット構築

時間的なズレがふくまれた演奏データは入手が困難であり，大量に収集することが困難である．そのため本研究では，疑似的なズレを含んだデータを作成し，学習データとして使用する．一方で，実際の演奏データでの評価も必要であり，小規模な演奏データの収録を行い，評価データとして用いる．初心者を模擬したデータを作成する方法として，本研究では一定のリズムで演奏することができない初心者の演奏を模擬する操作を演奏データに行う．様々な楽器編成の演奏データからギターとドラムの演奏データのみ抽出し，ギターの演奏データに対して，初心者を模擬する操作を行うことで，一定のリズムで演奏することができない初心者の演奏を模擬した演奏データを作成する．

4.1 模擬データ作成方法

RWC-MDB のポピュラー楽曲 [15] の MIDI データを使用し，ギタートラックに対して初心者の演奏を模擬するために，ギタートラックのイベントのタイミングをランダムにずらす操作を行った．この時，音響信号のずれの知覚に関する文献をもとにズレ幅を検討し，最小で 70 ms であるとした [16]．こうして作成された MIDI データから波形データを作成することで演奏タイミングのズレを含んだ演奏音データを作成した．また，MIDI データからギター，ドラムそれぞれについてのビートラベルを抽出し，音楽データに対応したラベルファイルを作成することで，データセットを構築した．その結果，89 曲，20,791 秒 (約 6 時間) のデータセットが構築された．

4.2 実演奏データ収録方法

20 代の学生 2 名の実験協力者の演奏を収録する．今回は，エレキギターを用いる．収録する楽曲として RWC-MDB のポピュラー楽曲のうち，10 曲を選別し，1 曲が 1，2 分程度になるように一部分を抽出した．なお，楽曲の選別は，ギター経験者へのアンケートを基に行った．演奏者は事前に楽譜を渡され，止まらず弾けるようになるまで練習し

た．収録の際，ギターの演奏音を USB オーディオインタフェース (Vox 社製 AMPLUG I/O) を介して PC に録音し，演奏者はドラムパートの演奏をヘッドホンで受聴しながら演奏した．

4.3 実演奏データに対するビート位置のラベリング

波形編集ソフトウエア Audacity を用いて収録された演奏音に対して手動でビートラベルの付与を行った．主にギター音のスペクトログラムと波形，演奏音のほか必要に応じて楽譜を参照した．なお，ラベルの付与は執筆者 1 名により行われた．ドラム演奏は MIDI から生成されており，ドラムパートのビートラベル列を得ることは容易である．ギター演奏のオンセット位置をスペクトルと波形，演奏音から決定し，ドラムパートのビート位置を参考に，ギターパートのビート位置に補正する作業を行った．

5. 評価実験

5.1 予備実験

他のビートトラッキング研究と異なり，本研究では演奏楽器をギターとドラムに限定している．これは曲の構成要素が限定されているためビートトラッキングの精度への何らかの影響があると考えられる．従って，本研究で構築したデータセットではなく，RWC-MDB のポピュラー楽曲を用いて，楽器を限定した演奏と、そうでない演奏にビートトラッキングを行い、その結果を比較する．なお，以降の評価実験において，実験時の特徴量の抽出には librosa[17] を用いる．また，特に記述がない場合，他のビートトラッキングの研究と同様に，評価実験において ± 70 ms の誤差基準 [18] を使用し，算出に音楽情報処理システムに対する評価方法が実装されたライブラリ mir eval[19] を用いる．

本研究で作成したズレを含んだ演奏データではなく RWC-MDB のポピュラー楽曲 100 曲からギターとドラムトラックを MIDI データ上で抽出した波形化したデータ 89 曲に対してビートトラッキングを行い，その影響の有無を確認する．この時，除外した 11 曲はデータセットを構築した際に除外した 11 曲と同じである．

ギターとドラムを抽出したデータに対して，ビートトラッキングを行った結果，F 値が 0.760 であり，すべての楽器の演奏データに対して同じ実験条件で実験を行った結果，F 値が 0.772 であった．t 検定を行った結果，p 値が 0.740 となり，有意水準 0.05 として有意差無しという結果になった．従って，これらの結果から演奏楽器を限定したことによるビートトラッキングの精度への影響は無いことが確認された．

5.2 実験 1: 初心者を模擬したデータに対する実験

模擬データに対する提案手法の有効性の検証を行う．従来技術の組み合わせによる比較対象手法として，音源分離と通常のビートトラッキング手法を組み合わせた手法を用いる．このような手法では，ずれを含んだ音源を個々のパートの音源に分離することにより，ずれの影響を排除し，通常のビートトラッキング問題に落とし込めるため，従来技術を組み合わせた実現方法として有用な手法であると考えられる．音源分離手法として Harmonic-percussive source separation（HPSS）と，ニューラルネットワークの一つである U-Net を用いており，原稿執筆時点では最新の音源分離手法である Spleeter を用いる（比較手法 1 および 2）．参考指標として，理想的な音源分離の場合として重畳する前の分離音を用いた手法（比較手法 3）においても実験を行う．

表 1 模擬データに対する実験結果 (F 値)

	ドラム	ギター
提案手法	0.754	0.603
比較手法 1 (HPSS)	0.760	0.578
比較手法 2 (Spleeter)	0.770	0.428
比較手法 3 (理想的な分離音)	0.764	0.404

表 1 に示した実験結果について，楽器ごとに着目してそれぞれ考察する．ドラムに関しては，提案手法と比較手法に大きな差は確認できなかった．一方でギターに関しては．比較手法と比べると提案手法が最も高い精度であり，ギターのソロ演奏では適切に学習できないが，入力に混合音を使用し，ドラムとギターのビートラベルを用いて同時に学習を行うことで，適切な学習ができることが示され，提案手法の有効性を確認することができた．

5.3 実験 2: 実演奏データに対する実験

実演奏データに対する提案手法の有効性の検証と模擬データの妥当性の検証を行う．この実験では収録した実演奏データを評価データとして用い，学習データとして作成した模擬データ用いる．実演奏データに対する実験結果を表 2 に示す．模擬データと同様に，ドラムとギターを比較すると，ドラムの方が高い精度が示されている．模擬データへの実験結果と比較して精度が低いため，今回収録した実演奏データではビートトラッキングの難易度が比較的高いものであるほか，モデルの学習に用いた疑似的な演奏ずれを含んだデータとの乖離があった可能性がある．

そのため，モデル学習時の検証データに実演奏データの半分を加えて学習し，残りの実演奏データに対して評価を行う．表 3 に演奏者 1 に対する実験結果を示す．表 2 の結果と比較すると，全ての評価尺度について，精度が改善している．このことから，実演奏データに対するビートトラッキングが困難であったことよりも，模擬データと実演奏データとの間の乖離が大きすぎることに問題があると考えられる．

表 2 実演奏データに対する実験結果 (F 値)

	ドラム	ギター
演奏者 1	0.711	0.531
演奏者 2	0.754	0.728

表 3 検証データに実演奏データを用いた実験結果 (F 値)

	ドラム	ギター
演奏者 1	0.778	0.599

6. まとめと今後の課題

本研究では，楽器ごとに一つのビートラベル列が存在し，複数のビートラベル列を持つ音楽に対して，混合音に対して複数パートのビート位置を同時にトラッキングするアプローチの下でマルチパートビートトラッキングのための新たな手法を提案した．提案手法に対して，従来手法を組み合わせた方法と比較することで有効性を示した．

今後の課題として，ズレに対する再現率を改善することが挙げられる．また，本研究の問題設定は，2 種類の楽器が演奏しており，1 種類の楽器のみ演奏タイミングがずれている演奏を対象としているが，双方の楽器の演奏タイミングがずれている演奏や，楽器の種類が増えた演奏を対象とすることも挙げられる．

評価実験の結果より，本研究で作成した模擬データは実演奏データと演奏の傾向に乖離があることが示唆された．実演奏データに対してビートラベルを付与した際，実演奏データを分析した結果，演奏タイミングのズレは局所的には一定にずれる傾向が確認された．従って，一定の範囲でランダムなズレを付与するのではなく，局所的には一定のテンポを持っている模擬データを作成することも今後の課題として挙げられる．

また，単独の評価者により付けられた実演奏データへのビートラベルの信頼性の検証も必要である．信頼性を上げる方法として，複数人でラベル付与を行う方法や，オンセット検出を行いビートか否かの選択をする半自動化によって信頼性やラベリングの揺れを低減させる方法などが考えられる．

謝辞

Jakub Gałka 博士 (ポーランド・AGH 科学技術大学) から提案手法の実装に関して有益な助言を受けた．

参考文献

[1] MIREX Wiki, https://www.music-ir.org/mirex/wiki/MIREX_HOME.
[2] Goto, M. and Muraoka, Y.: A beat tracking system for acoustic signals of music, *Proceedings of the second ACM international conference on Multimedia*, pp. 365–372 (1994).
[3] Elowsson, A.: Beat Tracking with a Cepstroid Invariant Neural Network, *17th International Society for Music Information Retrieval Conference*, pp. 351–357 (2016).
[4] Böck, S. and Schedl, M.: Enhanced Beat Tracking with Context-Aware Neural Networks, *Proceedings of the 14th International Conference on Digital Audio Effects, DAFx 2011*, pp. 135–139 (2011).
[5] 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*, pp. 603–608 (2014).
[6] Böck, S., Krebs, F. and Widmer, G.: Joint Beat and Downbeat Tracking with Recurrent Neural Networks, *International Society for Music Information Retrieval Conference*, pp. 255–261 (2016).
[7] Krebs, F., Böck, S. S., Krebs, F. and Widmer, G.: Rhythmic pattern modeling for beat and downbeat tracking in musical audio, *International Society for Music Information Retrieval Conference*, pp. 227–232 (2013).
[8] Holzapfel, A., Davies, M. E. P., Zapata, J. R., Oliveira, J. L. and Gouyon, F.: Selective Sampling for Beat Tracking Evaluation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 9, pp. 2539–2548 (online), DOI: 10.1109/TASL.2012.2205244 (2012).
[9] Fiocchi, D., Buccoli, M., Zanoni, M., Antonacci, F. and Sarti, A.: Beat Tracking using Recurrent Neural Network: A Transfer Learning Approach, *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1915–1919 (online), DOI: 10.23919/EUSIPCO.2018.8553059 (2018).
[10] Fitzgerald, D.: Harmonic/Percussive Separation using Median Filtering, *13th International Conference on Digital Audio Effects (DAFx-10)*, pp. 1–4 (2010).
[11] Hennequin, R., Khlif, A., Voituret, F. and Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models, *Journal of Open Source Software*, Vol. 5, No. 50, p. 2154 (online), DOI: 10.21105/joss.02154 (2020).
[12] Krebs, F., Böck, S. and Widmer, G.: An Efficient State-Space Model for Joint Tempo and Meter Tracking, *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 72–78 (2015).
[13] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F. and Widmer, G.: madmom: a new Python Audio and Music Signal Processing Library, *CoRR*, Vol. abs/1605.07008 (online), available from ⟨http://arxiv.org/abs/1605.07008⟩ (2016).
[14] 福谷和貴，酒向慎司：演奏タイミングのズレを含む混合音に対するマルチラベルビートトラッキング，情報処理学会研究報告音楽情報科学 (MUS)，Vol. 2020-MUS-129, No. 2, pp. 1–4 (2020).
[15] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases, *Proceedings of the International Society for Music Information Retrieval Conference*, Vol. 2, pp. 287–288 (2002).
[16] Dixon, S.: Automatic extraction of tempo and beat from expressive performances, Vol. 30, No. 1, pp. 39–58 (2001).
[17] McFee, B., Raffel, C., Liang, D., Ellis, D. P., Matt McVicar, E. B. and Nieto, O.: librosa: Audio and music signal analysis in python., *Proceedings of the 14th python in science conference*, pp. 18–25 (2015).
[18] Davies, M., Degara Quintela, N. and Plumbley, M.: Evaluation Methods for Musical Audio Beat Tracking Algorithms, *Technical Report C4DM-TR-09-06* (2009).
[19] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Ellis, D. L. and PW., D.: mir eval: A transparent implementation of common mir metrics, *Proceedings of the 15th International Society for Music Information Retrieval Conference* (2012).