Vol.2019-MUS-122 No.27
Vol.2019-EC-51 No.27
2019/2/23

位相情報を考慮したRNNによるドラム自動採譜

河田 洋人[1,a)]　保利 武志[1,b)]　中村 和幸[1,c)]

概要：音楽音響信号に対する自動採譜は音楽情報処理における主要なタスクである．特にドラム演奏の自動採譜はテンポ，拍位置の推定といったリズムに関わるメタ情報抽出のための重要な要素技術である．近年のドラム自動採譜問題ではRNNをはじめとした時系列情報からの特徴量抽出に基づいた手法が良い精度を上げており，研究者達によって様々な手法が議論されている．従来のRNNによる実装では周波数領域におけるパワースペクトルや対数パワースペクトルが特徴量として用いられてきたが，本稿では位相情報を特徴量に追加したRNNによるドラム自動採譜手法の提案を行う．評価実験の結果，位相情報を特徴量に加えることでドラムの打点推定精度は向上し，位相情報のドラム採譜における特徴量としての有効性が示唆された．

Automatic drum transcription based on RNN considering phase information

HIROTO KAWATA[1,a)]　TAKESHI HORI[1,b)]　KAZUYUKI NAKAMURA[1,c)]

Abstract: Automatic drum transcription (ADT) is a task of retrieving symbolic representation of drum instruments from monaural drum solo recordings. In recent years, recurrent neural network (RNN)-based approach have achieved the highest evaluation accuracies, therefore, various architectures have been discussed by many researchers. In the conventional RNN implementations, power spectrum in the frequency domain have been used as the input feature. In contrast to previous researches, for the purpose of improving the evaluation accuracy, we propose a new RNN architecture with phase information added to input layer. As a result of the evaluation experiment, by adding the phase information to the input, the accuracy of estimating the hitting time of drum instruments considerably increased, and the effectiveness of phase information is demonstrated.

1. はじめに

自動採譜は音楽情報科学における主要なタスクである．演奏者がスコアを見ながら演奏を行うのに対し，自動採譜は，この逆問題として，音楽音響信号から演奏の記号情報を抽出することを目的としている．情報技術の発達に伴い，教育，音楽制作，音楽情報検索といった様々な場面への自動採譜の貢献が期待され，高精度の自動採譜技術の必要性は高い．

ドラム自動採譜は自動採譜技術のドラム演奏への適用であり，ドラム演奏の音響信号から打楽器の打点時刻を含んだ記号表現を抽出することを目的としている．特に西洋音楽においてドラムスはリズムを構成する主要なパートであるため，ドラム自動採譜はテンポ，拍位置の推定といったリズムに関わるメタ情報抽出のための重要な要素技術となっている．また，楽曲構造 [1] や音楽ジャンル [2] との関連や，楽曲に内在するグルーヴとの関係性 [3], [4] も報告されており，構造解析，ジャンル分類，音響心理的な立場からも極めて重要な問題である．

ドラム自動採譜は，これまでに多くの研究者たちによって議論が行われてきたが，特に近年は時系列情報からの特徴量抽出に基づき，各楽器のアクティベーションへの射影を行う，アクティベーションベースのアプローチが最も良い精度を誇っている [5]，非負値行列因子分解 (Non-negative

[1] 明治大学 大学院先端数理科学研究科
〒 164-8525 東京都中野区 4-21-1
[a)] cs181004@meiji.ac.jp
[b)] hori@meiji.ac.jp
[c)] knaka@meiji.ac.jp

Vol.2019-MUS-122 No.27
Vol.2019-EC-51 No.27
2019/2/23

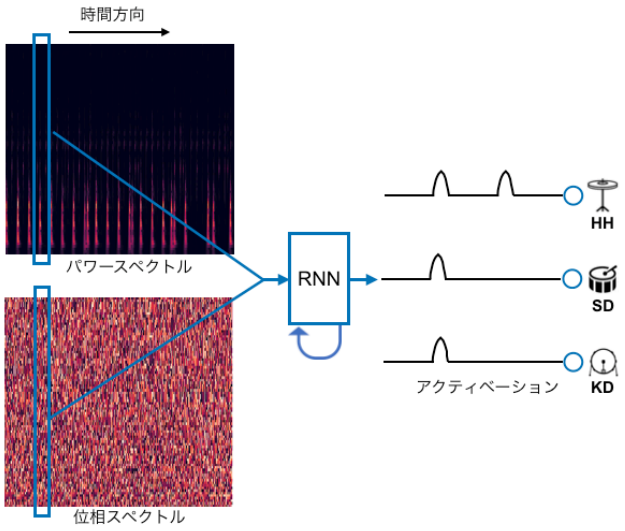

図 1　位相を考慮した RNN によるハイハット，スネアドラム，キックドラムのアクティベーション推定

matrix factorization, NMF) を用いた手法とニューラルネット (Neural networks, NN) を用いた手法はそれらの代表的である．

NMF は非負値行列を異なる 2 つの低ランクな非負値行列の積に分解するアルゴリズムで，各楽器のスペクトルパタンと対応するアクティベーションに分解する．Dittmar ら [6] は NMF をモノラルドラム音源に適用しハイハット (HH)，スネアドラム (SD)，キックドラム (KD) の 3 つのクラスへ分類を行い，95 %以上の F 値を得た．

ニューラルネットワークはノードと重み付きの指向性グラフから構成される．近年，複雑なネットワーク構成の学習を可能にする最適化手法の発見 [7] により，利用される機会の多い機械学習手法である．Vogl ら [8] は Recurrent Neural Networks (RNN) を用いて時系列を考慮したドラム演奏のアクティベーション推定を行った．その推定精度は NMF を上回る 98.2 %を誇り，RNN のドラム自動採譜への有効性が示された．他にも LSTM，GRU，畳み込みを用いたもの [9], [10], [11] が報告されており，RNN 手法の有効性は高い．しかしながら，RNN の入力には複素フーリエ係数のパワーが用いられ，位相情報は入力の際に省かれてきた．

本来位相はオンセット時刻周辺にて連続性を保つという性質を持ち，新たな楽器の打点時には，その連続性が崩れることから位相情報のオンセット検出への有効性が示されている [12]，そこで，本報告では，近年主流となっている RNN 手法の更なる精度向上を目的として，位相情報を入力に考慮した RNN によるドラム自動採譜手法を提案し，位相情報の有効性について述べる．

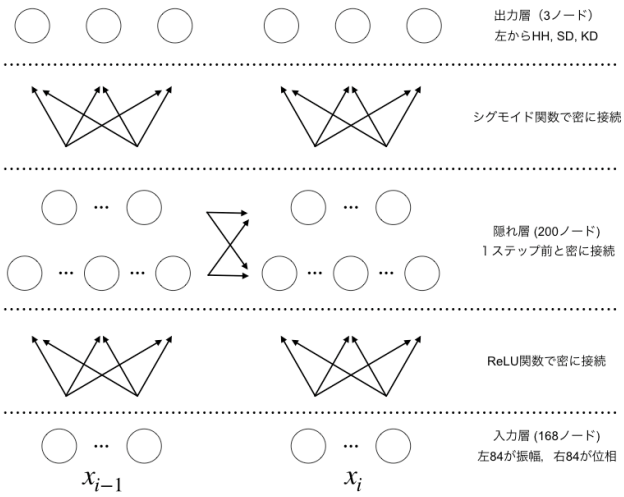

図 2　RNN のネットワーク構成

2. ドラム自動採譜問題の定式化

2.1 ドラム自動採譜問題の定義

本報告では，ドラム自動採譜問題をハイハット，スネアドラム，キックドラムから構成されたモノラルドラム音源からの各楽器の打点時刻推定問題として扱う．ハイハット，スネアドラム，キックドラムはドラム演奏を構成する最も基本的な打楽器であり，タムやシンバルなどのその他の打楽器は演奏の装飾として用いられることが多い．この理由から，推定される対象はこれらの 3 つに限定されることが多く，本研究においてもそれに従う．

2.2 位相を考慮した RNN によるドラム楽器のアクティベーション推定

ドラム自動採譜を行うためには、複数の打楽器を含む複合音源から各楽器の打点時刻を算出する必要がある．しかし，ドラム演奏の際には，複数楽器を同時に打点するため，単純なオンセット検出では重複した楽器の打点時刻推定は困難である．そこで，ドラム音源の音源分離を行い，各楽器のアクティベーションを算出し，多重音源のオンセット検出を可能にする必要がある．

本報告では，出力がハイハット，スネアドラム，キックドラムのアクティベーションとなるような RNN をデータから学習することで、与えられた音源に対する各楽器のアクティベーションの算出を実現する．RNN は隠れ層に再帰構造を持つ深層学習の一種で，再帰構造によりシークエンスデータに有効なネットワークとなっており，自然言語処理などにおいて高い精度を発揮している．ドラム自動採譜への応用は [8] によって報告され，ドラム楽器の打点時刻推定を従来手法を上回る精度で行った．しかしながら，上記のモデルにおける入力は複素フーリエ係数のパワーのみを用いており，位相情報は入力の際に省かれていた．

本来，位相はオンセット時刻周辺で連続性を保つという

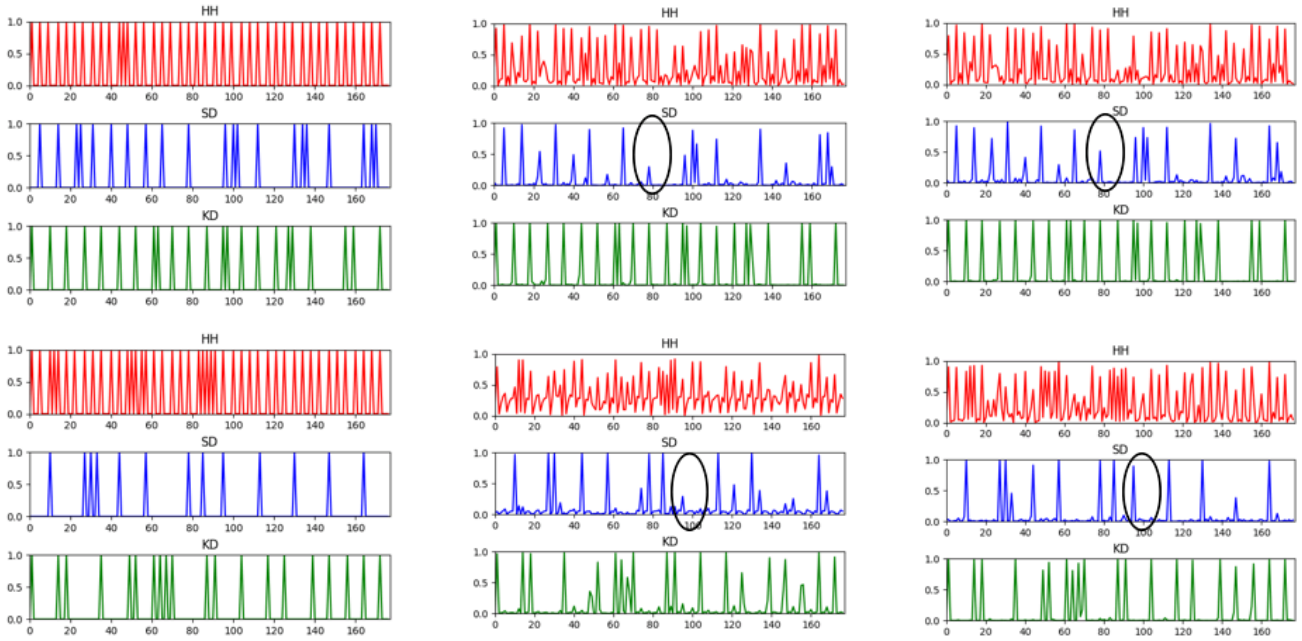

図 3 上段，下段それぞれ別の音源に対して，左から正解ラベル，従来手法による推定アクティベーション，提案手法による推定アクティベーション

性質を持ち [12]，位相を入力として持つことで，新たな楽器の打点時にはその連続性が途絶え，ネットワークがその状況を学習し，推定精度の向上が見込まれる．これらの理由から，本報告では通常の振幅スペクトルに位相情報を特徴量として追加したネットワーク (図 1) を提案する．

2.3 ネットワーク構成

図 2 は本報告で用いる RNN のネットワーク構成を表す図である．入力層は 168 個のノードを持ち，7 オクターヴ分の定 Q 変換により得られる 84 次元複素フーリエ係数の振幅と位相から構成される．隠れ層は 200 個のノードから構成され，1 ステップ前との接続を持つ．出力層はハイハット，スネアドラム，キックドラムのアクティベーションを表す 3 個のノードを持つ．入力層の活性化関数には，ReLU 関数が，隠れ層の活性化関数にはシグモイド関数が用いられ，全てのグラフは密に接続される．入力層以外のネットワーク構成は Vogl ら [8] を参考にした．

3. 評価実験

3.1 実験の目的

これまでに述べた，位相情報を考慮した RNN により，ハイハット，スネアドラム，キックドラムのアクティベーション推定を行った．また，位相情報を入力に考慮しないモデルでもアクティベーション推定を行い，誤差の比較をすることで，位相情報の入力としての有効性を検証した．そのために，提案モデルとは別に，位相を除いた 84 ノードの入力層を持ち，その他のパラメータ，ネットワーク構成が全て同一のモデルを用意した．

3.2 学習，評価用のデータセット

ドラム自動採譜用のデータセット IDMT-SMT-Data 内のアコースティックドラム演奏音源，計 20 を用いて，学習と評価を行った．各音源では，ハイハット，スネアドラム，キックドラムが任意のタイミングで叩かれ，その打点時刻が別ファイルにラベル付けされている．音源はサンプリング周波数 44100Hz，16bit 解像度で録音がされており，録音時間の合計は 300 秒，アノテーションの合計は 1880 個であった．20 音源の内，18 音源を学習に用い，2 音源をハイハット，スネアドラム，キックドラムのアクティベーション推定に用いた．

3.3 音源の前処理

サンプリングレート 44100Hz，16bit 解像度のドラム演奏音源を，25600Hz にダウンサンプリングした．その信号に対して 7 オクターヴ分の定 Q 変換による特徴量抽出を行い，84 次元の複素フーリエ係数を算出し，その振幅と偏角を 168 次元の入力ベクトルとした．また，RNN への入力のために 100 ステップごとにフレーム化を行った．

3.4 学習

ハイハット，スネアドラム，キックドラムそれぞれのオンセット時刻が 1 と記述されたラベルを正解アクティベーションとして，音源の前処理にて作成したフレームと共にネットワークの学習を行った．パラメータの最適化には，

表 1　正解，推定アクティベーション間の RMSE の比較

	従来手法 (振幅のみ)	提案手法 (振幅、位相)
HH	0.203	0.202
SD	0.187	0.174
KD	0.0430	0.0387

表 2　正解，推定アクティベーション間の RMSE の比較

	従来手法 (振幅のみ)	提案手法 (振幅、位相)
HH	0.304	0.205
SD	0.130	0.0676
KD	0.107	0.0360

検証の際に最も精度が高かった RMSProp を用い，batch size=8, epoch=100 として学習を行った．

3.5　結果

図 2 は上段，下段ごとに，左から正解の打点時刻，従来モデル（振幅のみが入力）のアクティベーション推定結果，提案モデル（振幅，位相が入力）のアクティベーション推定結果をプロットした結果である．また，表 1，2 では，正解アクティベーションと推定アクティベーション間の Root mean squared error (RMSE) を計算することで従来手法と提案手法の予測精度を比較した．

図 3 において，テスト音源 1，2 共に従来手法と提案手法を比較すると全体的にノイズの軽減が見られ，一部アクティベションが鋭くなった．実際に，表 1，2 からはハイハット，スネアドラム，キックドラム全てにおいて推定誤差の減少が観測され，位相情報の有効性が示された．

4.　おわりに

4.1　結論

本報告では，ドラム演奏の音響信号に対する自動採譜手法の精度向上を目的とし，位相情報を考慮した RNN の提案を行った．2 つのテスト音源に対して，アクティベーション推定を行った結果，ノイズの軽減やピークが鋭くなることが観測された．また，RMSE の比較ではハイハット，スネアドラム，キックドラム全てにおいて誤差の減少が確認され，位相情報の入力としての有効性が示された．

今後の課題としては位相情報の有効性仮説を立証することである．現段階においては，推定精度の向上は確認できたものの，その具体的な仮説の検証を行うことができなかった．そのために，ドラム演奏の位相スペクトルの解析やモデルの入力のうちどこが有効であるかを調査し，精度向上の要因を示していきたい．

謝辞　本研究は JSPS 科研費 17H00749 の支援を受けた．

参考文献

[1] Kristoffer Jensen: “Multiple scale music segmentation using rhythm, timbre, and harmony”, EURASIP Journal on Applied Signal Processing, Vol.2007, PP.159-159, 2007

[2] Simon Dixon, Fabien gouyon, and Gerhard Widmer: “Towards Characterisation of Music via Rhythmic Patterns”, in ISMIR, 2004

[3] 宮丸 友輔, 江村 伯夫, and 山田 真司: “ポピュラ音楽のドラムス演奏におけるグルーヴ感の研究”, 日本音響学会誌, Vol. 73, No.10, pp.625-637, 2017

[4] 奥平 啓太, 平田 圭二, 片寄 晴弘: “ポップス系ドラム演奏の打点時刻及び音量とグルーヴ感の関連について”, 情報処理学会研究報告音楽情報科学, Vol.84, pp.21-26, 2004

[5] Chih-Wei Wu, Christian Dittmar, Carl Southall, Gerhard Widmer, Jason Hockman, Meinard Mller, and Alexander Lerch: “A Review of Automatic Drum Transcription”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.26, No.9, pp.1457-1483, 2018

[6] Christian Dittmar, Daniel Grtner: “Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition”, in DAFx, pp.187-104, 2014

[7] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh: “A Fast Learning Algorithm for Deep Belief Nets”, Neural Computation, Vol.18, No.7, pp.1527-1554, 2006

[8] Richard Vogl, Matthias Dorfer, and Peter Knees: “Recurrent neural networks for drum transcription”, in ISMIR, 2016

[9] Richard Vogl, Matthias Dorfer, Peter Knees: “Drum transcription from polyphonic music with recurrent neural networks”, in ICASSP, 2017

[10] Carl Southall, Ryan Stables, and Jason Hockman: “Automatic drum transcription using bi-directional recurrent neural networks”, in ISMIR, 2016

[11] Carl Southall, Ryan Stables, and Jason Hockman: “Automatic drum transcription for polyphonic recordings using soft attention mechanisms and convolutional neural networks”, in ISMIR, 2017

[12] Juan P. Bello, Chris Duxbury, Mike Davies, and Mark Sandler “On the use of phase and energy for musical onset detection in the complex domain”, IEEE Signal Processing Letters, Vol.11, No. 6, 2004