1V-05

無音動画に着目したドラム演奏の自動採譜

井村　悠斗†　　上田　芳弘‡　　坂本　一磨‡

公立小松大学大学院サステイナブルシステム科学研究科†　公立小松大学生産システム科学部‡

１．はじめに

ドラム演奏の自動採譜とは，主にドラム演奏の音響を楽譜等の記号表現に変換するタスクである．しかし，音響のみを入力とする手法では，ポリフォニックな演奏の採譜や，背景雑音が存在する環境下で録音された演奏の採譜が困難となる場合が多い．多用な収録環境に適応し，自動採譜の汎用性を向上させるためには，音響情報のみではなく，視覚情報を用いることが有効であると考えられる．本稿では，電子ドラムを用いたデータ収集と，LTC（Linear Timecode）を用いた自動アノテーション，深層学習モデル ECO Lite[1]による自動採譜の実装を提案し，視覚情報のみによる自動採譜の可能性を検証する．

２．提案手法

２．１　データ収集

データ収集の手順を図 1 と以下に示す．

(1) DAW（作曲ソフトウェア）の収録開始

DAW（Digital Audio Workstation）の音響トラックにあらかじめ LTC を用意（以下，LTC トラック）しておき，DAW の収録を開始する．収録開始と同時に LTC が再生され，ビデオカメラに出力する．MIDI トラックには，電子ドラムから出力された MIDI データが記録される．LTC とは，短波と長波により 0 と 1 を表現することで，フレーム単位の時刻情報を示す音響データである．

(2) ビデオカメラの収録開始

撮影環境の様子の録画を開始する．DAW から出力されている，この時点からの LTC も同時に記録が開始される．

(3) 演奏の開始

演奏者が電子ドラムの演奏を開始する．電子ドラムから出力される MIDI データは DAW に記録し，演奏の様子はビデオカメラに記録する．

２．２　自動アノテーション

電子ドラムと DAW を用いて収集した MIDI データから，楽器が叩打されたタイミングか否か（1：Onset or 0：Silence）を示す学習用正解ラベルを楽器毎に作成する．手順は以下の通りである．

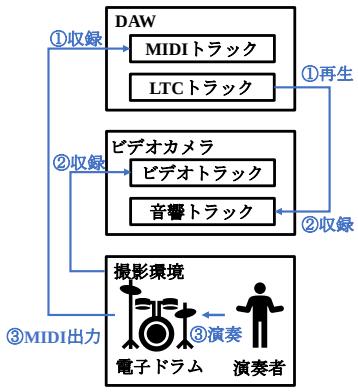

図 1　データ収集

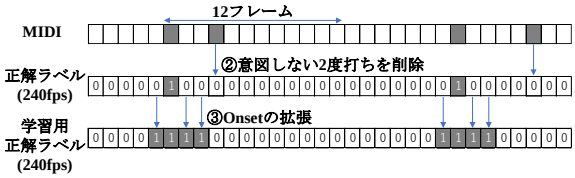

図 2　MIDI から正解ラベルへの変換

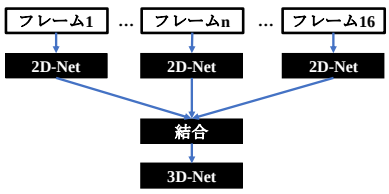

図 3　ECO Lite の構造

(1) MIDI と映像データの同期

LTC は，通常 240fps の時間分解能には対応していない．そのため，LTC の波形を読み取り，DAW とビデオカメラの LTC を照合するプログラムを作成し，収録開始時刻の差分を取得することで，MIDI と映像データの時間軸を同期させる．

(2) 意図しない 2 度打ちの削除

ダウンストローク奏法時，スティックのバウンドにより，大音の直後に演奏者の意図しない小音が出現する場合がある．収録した MIDI を視聴した際，小音は大音によってかき消され，聞き取ることが困難である．そのため，この小音（以下，意図しない 2 度打ち）は採譜対象外として，12 フレーム（≒54ms）以内に 2 つの Onset が検出された場合，図 2 に示すように削除する．

(3) Onset の拡張

アノテーションの誤差を吸収するため，Onset

Automatic Drum Transcription Focusing on Silent Video

†Yuto Imura・Graduate School of Sustainable Systems Science, Komatsu University

‡Yoshihiro Ueda and Kazuma Sakamoto・Faculty of Production Systems Engineering and Sciences, Komatsu University

の前 1 フレーム，後 2 フレームの Silence を Onset とした（図 2）．

２．３　動画認識モデルの学習

ECO Lite は，動画認識のための深層学習モデルである．図 3 に ECO Lite の構造を示す．動画からサンプリングした複数枚のフレームそれぞれに対して 2D-Net（BN-Inception[2]）を適用した後，2D-Net の出力を結合し，3D-Net（3D-ResNet18[3]）に入力している．この構造により，高い性能と処理速度の向上を両立している．

本稿では，Kinetics-400[4]により事前学習された ECO Lite を演奏動画によって，ファインチューニングした．ドラム採譜モデルは，16 枚のフレーム$f_{i-32}, f_{i-28}, f_{i-24}, \cdots, f_i, \cdots, f_{i+28}$を入力とし，注目するフレーム$f_i$に対して 2 値分類（Onset or Silence）を行う．

３．評価実験

３．１　撮影条件

iPhone14 を用いて，240fps の動画を撮影した．撮影視点は，演奏者の右後方と頭上（図 4）とし，ハイハット（HH），スネアドラム（SD），バスドラム（BD）の打面が映るよう撮影した．演奏者は 1 名とし，HH，SD，BD のみを用いた単純なフレーズの動画を，17 データ用意した．

３．２　データセット

用意した 17 の動画から学習データ，検証データ，評価データを構成する．それぞれのデータが含む Onset 数は，学習データが HH：178, SD：127, BD：118，検証データが HH：217, SD：70, BD：115，評価データが HH：347：SD：135, KD：142 である．全フレームの内，Onset は疎であるため，学習データと検証データは Silence をダウンサンプリングした．

３．３　評価

学習データによってファインチューニングした 3 つ（HH，SD，BD）のドラム採譜モデルに対して評価データを入力した結果，図 5 に示す通り，推論結果（240fps）は，Onset 周辺の Silence を Onset であると誤分類していた．そのため本稿では，実用環境に近い評価を行うため，時間分解能の変換（240fps から 30fps），連続したラベル 1 の修正，許容検出誤差 30fps（≒33ms）を適用した評価用正解ラベルを用い，適合率（P），再現率（R），F 値（F）による評価を行う．

４．結果と考察

撮影視点毎の評価結果を表 1 に示す．なお，Silence の結果は全て 1.00 を示したため割愛する．

評価データ全体（右後方＋頭上）の精度は，各楽器に共通して，高い再現率を示した一方で，適合率が比較的に低い値を示した．このことから，正解ラベルに含まれるほぼ全ての Onset を検出することができたが，そのうち HH, SD は 19%, BD は 6%が誤検出であったことがわかる．

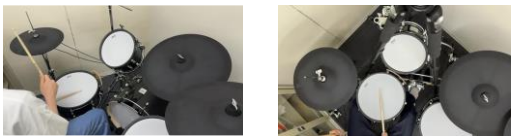

図 4　撮影視点（右後方，頭上）

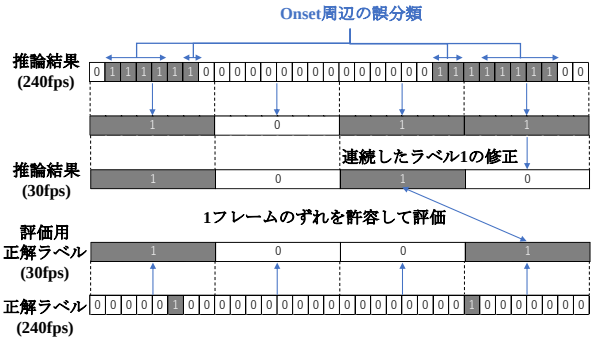

図 5　評価方法

表 1　評価結果

	右後方＋頭上			右後方			頭上		
	HH	SD	BD	HH	SD	BD	HH	SD	BD
P	0.81	0.81	0.94	0.83	0.73	0.98	0.78	0.91	0.89
R	1.00	0.94	0.99	0.99	0.89	1.00	1.00	1.00	0.98
F	0.89	0.87	0.97	0.91	0.80	0.99	0.87	0.95	0.93

HH は各視点で共通して，スティック先端の軌道が，HH 打面とオクルージョンしながら通過するタイミングで Onset の誤検出が発生する傾向があった．この結果から，画像の奥行を認識することが困難であったものと考えられる．

SD は右後方視点の場合，2.2 節で述べた意図しない 2 度打ちを Onset として誤検出する傾向があった．一方で，頭上視点の場合は意図しない 2 度打ちによる誤検出は少数であった．

５．おわりに

本稿では，HH, SD, BD のみの条件下での，叩打された楽器の特定とタイミングの検出は達成したが，検出された Onset は誤検出が目立つ結果となった．今後はデータセットの増強等により，汎用性と精度の向上を目指す必要がある．

参考文献

[1] Zolfaghari, M., Singh, K., and Brox, T. : ECO: Efficient Convolutional Network for Online Video Understanding, *ECCV 2018*, Vol. 11206, pp. 713-730, (2018).

[2] Ioffe, S., Szegedy, C. : Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, *ICML 2015*, Vol. 37, pp. 448-456, (2015).

[3] Tran, D., Ray, J., Shou, Z., Chang, S., and Paluri, M. : ConvNet Architecture Search for Spatiotemporal Feature Learning, *arXiv preprint arXiv:1708.05038*, (2017).

[4] Kay, W., et al. : The Kinetics Human Action Video Dataset, *arXiv preprint arXiv:1705.06950*, (2017).