5D-02

ドラムグルーヴ解析における LSTM 変分オートエンコーダ利用の検討

松川　瞬†　　竹沢　恵†　　稲垣　潤†　　真田　博文†

北海道科学大学†

1　はじめに

近年の音楽体験においては，映像と音楽のリズムが合致して身体的高揚感＝グルーヴを得ることが重要となる．しかし，視覚と聴覚の相互作用に関する研究は多いが，リズムの要となるドラムのグルーヴについて，音楽要素(奏者に依存しない要素)と演奏特性(奏者に依存する要素)の両方を踏まえ定量的に解明する研究は殆どない．

本研究では，楽曲データを入力として，音楽要素と演奏特性を同時かつ定量的に表現する機械学習モデルを構築し，ドラムグルーヴをデータドリブンに解析することを目標とする．本稿では，LSTM (Long Short-Term Memory)[1]の中間層出力をグルーヴの特徴と見做し，変分オートエンコーダ(Variational AutoEncoder，VAE)[2,3]により中間層出力分布(LSTM ユニット出力の隠れ状態分布)を 2 次元正規分布で表現する．その分布から，楽曲のグルーヴ感，特に演奏特性の差異の抽出・再構成を行い，定量的な分析を試みる．

2　グルーヴ要素の定義

ある演奏者p，あるセクションs(イントロ，サビといった楽曲内の区分け)におけるグルーヴが

$$\boldsymbol{gr}_{p,s} = \{notation, module, tone, dynamics, nuance\}$$

で構成されると定める．ここで，notation は音譜の配置，module はリズムパターン，tone は音色，dynamics はダイナミクス(強弱)，nuance はニュアンス(譜面とのずれ)を表す．このうち，notation・module・tone が音楽要素，dynamics・nuance が演奏特性である．なお，演奏者pの持つグルーヴは，セクションs内で不変とする．

本稿では，演奏特性の dynamics(特にゴーストノーツと呼ばれる「聞こえるか聞こえないか」と言う程度の弱音)，nuance(特にルーズ／タイトと表現されるタイミングズレ)に着目し，それらを含む演奏音源と，音の大きさ・タイミングが一定の打込音源との差異抽出を行う．それらは楽曲のグルーヴに大きく影響すると言われている[4]．

3　LSTM 変分オートエンコーダモデル

楽曲特徴の学習には，時系列データ解析に有効である LSTM を利用する．LSTM は，ニューラルネットワークにおける中間層のノードを図 1 に示すメモリーユニット(Memory Unit)に置き換えたモデルで，通常の入力受容部(Input)のほか，入力・忘却・出力ゲート(Input/Forget/Output gate)，過去の出力を再帰するメモリーセル(Memory cell)を持つ．各ゲートでの入出力を一般化して記述すると，時刻tにおける入力$\boldsymbol{x}_t$，再帰される値を$\boldsymbol{z}_{t-1}$，入力部の重さ$\boldsymbol{u}_{in}$，再帰部の重さ$\boldsymbol{u}_{re}$として

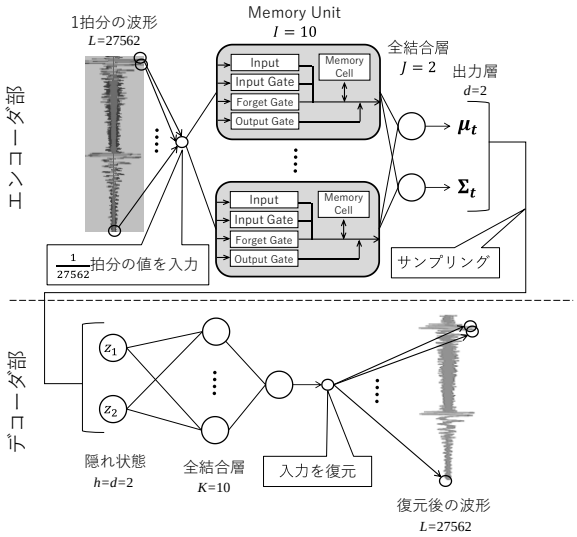

図 1　モデル概要図

$$g(\boldsymbol{x}) = f\left(\sum_t \boldsymbol{u}_{in}\boldsymbol{x}_t + \boldsymbol{u}_{re}\boldsymbol{z}_{t-1})\right) \quad (1)$$

となる(バイアスの記述は割愛)．fは活性化関数である．再帰値$\boldsymbol{z}_{t-1}$でモデルの過去状態を表現しつつ，忘却ゲートを通すことで，長期的な時間依存性を記憶できる．

本稿では，この LSTM の出力値を対象として，変分オートエンコーダにより隠れ状態の分布(正規分布)を推定する．変分オートエンコーダは，隠れ状態$\boldsymbol{z}$の分布$P(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$を推定するエンコーダ部分(式 2)と，その分布からサンプリングした値を基に出力を決定するデコーダ部分(式 3)に分かれる生成モデルの一種である(図 1)．

$$\boldsymbol{\mu}(\boldsymbol{x_t}),\boldsymbol{\Sigma}(\boldsymbol{x_t}) = f\left(\sum_i \boldsymbol{w}_i \cdot g_i(\boldsymbol{x_t})\right) \quad (2)$$

$$h(\boldsymbol{z_t}) = f\left(\sum_k \boldsymbol{v}_k \cdot \boldsymbol{z_t}\right), \boldsymbol{z_t} \sim N\big(\boldsymbol{\mu}(\boldsymbol{x_t}),\boldsymbol{\Sigma}(\boldsymbol{x_t})\big) \quad (3)$$

生成モデルとしては，学習後はデコーダ部分のみを使い，$\boldsymbol{z} \sim N(\boldsymbol{0},\boldsymbol{I})$としてサンプリングした値をデコーダに入力するが，本稿ではエンコーダ部分が持つ隠れ状態の分布表現能力，すなわちサンプリング元の分布に着目する．

A Study of Using LSTM Variational Autoencoders in Drum Groove Analysis

†Shun Matsukawa, Megumi Takezawa, Jun Inagaki and Hirofumi Sanada

†Hokkaido University of Science

4　カルバックライブラー情報量による隠れ状態分布の差異抽出

時刻tにおける演奏音源と打込音源との差異は，多変量正規分布のカルバックライブラー情報量D_{KL_t}で求める．

$$D_{KL_t} = \frac{1}{2}\left[\log\frac{\left|\boldsymbol{\Sigma}_{2_t}\right|}{\left|\boldsymbol{\Sigma}_{1_t}\right|} + tr\left(\boldsymbol{\Sigma}_{2_t}{}^{-1}\boldsymbol{\Sigma}_{1_t}\right) + \left(\boldsymbol{\mu}_{2_t} - \boldsymbol{\mu}_{1_t}\right)^T \Sigma_{2_t}{}^{-1}\left(\boldsymbol{\mu}_{2_t} - \boldsymbol{\mu}_{1_t}\right) - d\right] \quad (4)$$

ここで$\boldsymbol{\mu}_{1_t}, \boldsymbol{\Sigma}_{1_t}$は時刻$t$における演奏音源の隠れ状態の平均・共分散，$\boldsymbol{\mu}_{2_t}, \boldsymbol{\Sigma}_{2_t}$は打込音源の隠れ状態の平均・共分散，$d$は分布の次元数を表す．この数値は打込音源が演奏音源に変化する際の情報量の増量の期待値を表すため，「演奏音源にあって打込音源にない」特徴，すなわちグルーヴ(演奏特性)の差異を表す値であると考える．

5　実験結果

本稿では Red Hot Chili Peppers の Dani California を差異抽出の対象楽曲とし，演奏音源としてイントロのセクション 40 拍を抜き出した．打込音源は手作業で作成し，tone 以外の音楽要素は一致するが演奏特性を含まないものにした．モデル構造は図 1 に示す通りで，入力時系列長は楽曲 1 拍分＝27562 とした．学習は演奏音源のみで行い，差異抽出には演奏・打込両音源を用いた．

抽出結果として，図２に，ある 1 拍分の演奏音源・打込音源の波形(図上・中)と D_{KL_t}のグラフ(図下)を示す．横軸が時刻を表している．波形前半はわずかに nuance 差がある通常の打音，後半は dynamics・ゴーストノーツを持つような波形となっている．

D_{KL_t}のグラフを見ると，前半の通常の打音において差が大きく出ているが，後半の dynamics・ゴーストノーツ部においては殆ど差が出ていない．

後半部分に着目すると，LSTM 変分オートエンコーダによる特徴の差異抽出が音源波形自体の差異にあまり依存しないこと，つまり抽出された差異の楽曲 tone 依存性が少ないことが言える．すなわち，2章で上げたグルーヴ定義における音楽要素の部分は，本モデルによりほぼ無視でき，演奏特性のみにおける差異の抽出が可能であると考えられ，変分オートエンコーダがグルーヴ抽出に有用であると言える．

前半部分に注目すると，演奏音源における打音の開始点が微妙に早く(前の拍に僅かに掛かっている)，その部分においてD_{KL_t}が非常に大きくなっている．よって，隠れ状態はわずかな nuance の違いに強く反応すると考えらえる．逆に後半のゴーストノーツ部分ではD_{KL_t}が非常に小さく，dynamics の特徴差異は取得しづらいと言える．

ゴーストノーツの技法は古くから存在しており，楽曲のグルーヴに無関係であるとは考えづらい．得られた隠れ状態分布を見ると，前半部分の平均値は大きく変動しているが，後半部分は変動が小さく残響部分と似た分布になっていたことから，通常打音とゴーストノーツとの振幅差が大きくゴーストノーツがノイズの様に処理されてしまっている可能性が考えられた．今後，ラッパのような減衰形になる打楽器の波形に上手く対応できるモデルで再検討する必要がある．

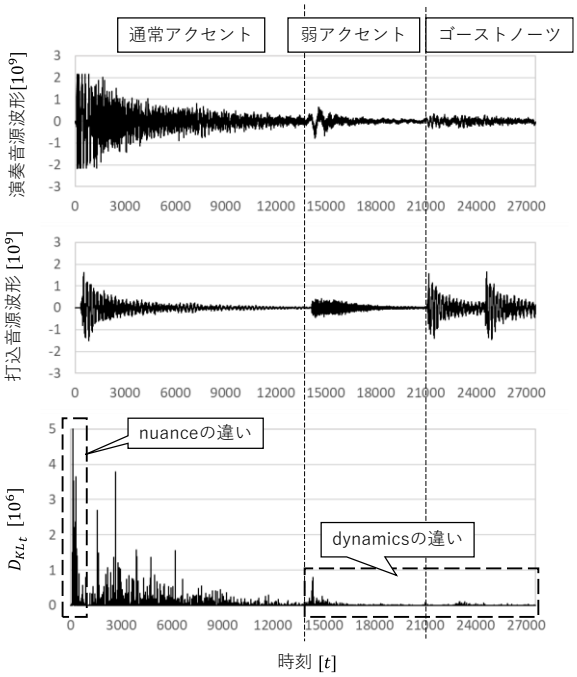

図２　音源間のカルバックライブラー情報量

6　結論

機械学習により楽曲データを入力として音楽要素と演奏特性を同時かつ定量的に表現するため，LSTM 変分オートエンコーダにより LSTM の中間層出力における隠れ状態分布を 2 次元正規分布で表現し，演奏音源・打込音源間の演奏特性の差異抽出・再構成を試みた．

音源間で隠れ状態分布のカルバックライブラー情報量を求めたところ，LSTM 変分オートエンコーダによる特徴の差異抽出は，音源波形自体の差異にあまり依存しないこと，nuance の違いに強く反応し dynamics の違いに殆ど反応しないことが分かった．

以上より，LSTM 変分オートエンコーダとカルバックライブラー情報量を用いた特徴の差異抽出・再構成により，グルーヴの定量的な分析が行えた．今後，dynamics の差異抽出が可能なモデルを構築し，抽出した差異と人間の感性との関係性を調査することで，本研究の目標を達成できると考える．

参考文献

[1] F.A. Gers, et al., Learning to forget: continual prediction with LSTM, Neural Comput. 12(10), pp. 2451-2471, 2000.

[2] D.P. Kingma, et al, Auto-Encoding Variational Bayes, arXiv:1312.6114, 2014.

[3] C. Doersch, Tutorial on Variational Autoencoders, arXiv:1606.05908, 2021.

[4] 奥平啓太他，ポップス系ドラム演奏の打点時刻及び音量とグルーブ感の関連について(第 2 報)，情報処理学会研究報告.MUS，vol.59，pp.27-32，2005.