E-057

リズムグルーヴ可視化における解釈可能性の改善に関する一検討
A Study on improving interpretability in rhythm groove visualization

松川 瞬† 松本 拓† 鈴木 昭弘† 荒澤 孔明† 松﨑 博季†
Shun Matsukawa Taku Matsumoto Akihiro Suzuki Koumei Arasawa Hiroki Matsuzaki

1. はじめに

近年の音楽体験として，楽曲を「聴く」だけでなく視覚・聴覚を刺激し「見る」「感じる」ことも非常に重要な点となっており[1]，映像の視覚効果と音楽を合致させる演出などによって音楽に「ノる」音楽サービスが様々に提供されている．そうして音楽から得られる身体的高揚感は「グルーヴ」と呼ばれる．

特に商業音楽（ジャズ・ロック・ポップス等）に着目したとき，グルーヴを生み出すには，楽曲の旋律もさながら打楽器・ドラムにおけるリズムの解析が非常に重要である．従来，記述統計量などのハンドクラフトな特徴量によるアプローチでドラムのグルーヴ解析を行う[2]ことが多かったが，ドラムの演奏には音楽要素（リズムパターン，譜面）と演奏特性（タイミングの微妙なズレやダイナミクス等，奏者の意図）が大きく関わっており，ハンドクラフトなアプローチではグルーヴ感の要となるそれら音楽要素・演奏特性を複合的に扱う事が非常に難しい．また演奏特性は複雑であり，適切な特徴の選択や定量化そのものが困難であるため，ハンドクラフトなアプローチでは限界がある．

これまで著者らは，LSTM（Long Short-Term Memory）[3]と変分オートエンコーダ（Variational Auto Encoder：VAE）[4]を組み合わせたモデルである LSTM 変分オートエンコーダ（LSTM-VAE）の中間層で得た確率分布間の情報量からリズム波形の特徴を定量的に取得・可視化することを試みてきた[5,6]．しかし，リズム波形に含まれるノイズの影響により，アクセントの弱い箇所における定量化・可視化が困難であり，ノイズに頑健かつ解釈可能性の高い中間層の表現方法が必要であった．

本研究では VAE に自己組織化マップ（SOM）を組み合わせた SOM-VAE[7]の考えを導入し，ノイズに頑健な特徴取得かつ可視化結果の解釈可能性の改善方法について検討する．

2. LSTM-VAE によるグルーヴ抽出

2.1 モデル概略

LSTM 変分オートエンコーダは，LSTM の出力を隠れ状態$\boldsymbol{z}$の分布のパラメータと見做すエンコーダ部分$f_\theta(\cdot)$と，その分布からサンプリングした隠れ状態の値を基に出力を決定するデコーダ部分$g_\phi(\cdot)$に分かれる．エンコーダ部とデコーダ部は，サンプリング層を通して繋がる．

入力データ$\boldsymbol{x} \in \mathbb{R}^d$はエンコーダ$f_\theta(\cdot)$により隠れ層の分布のパラメータの潜在空間$\boldsymbol{\eta} \in \mathbb{R}^m$に写像される（$\boldsymbol{\eta} = f_\theta(\boldsymbol{x})$）．隠れ状態はサンプリング$\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{\eta})$により得られた値となり，それがデコーダ$g_\phi(\cdot)$へ入力され，元の入力データを再現する（$\hat{\boldsymbol{x}} = g_\phi(\boldsymbol{z})$）．なお，隠れ状態の確率分布は通常ガウス分布$N(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\sigma})$が用いられる．

† 北海道科学大学 情報科学部 Hokkaido University of Science, Faculty of Information Science and Technology

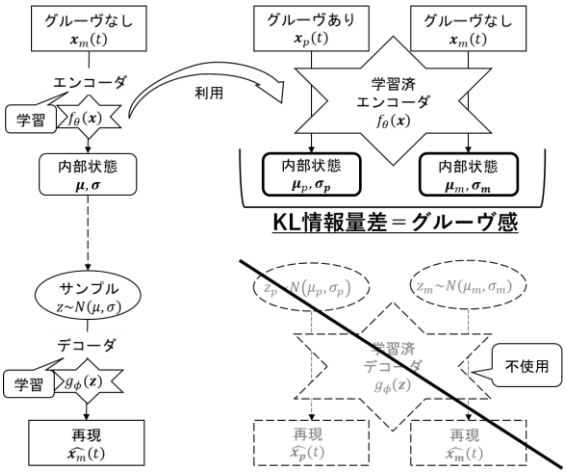

図 1 LSTM-VAE モデル概略図

変分オートエンコーダを組み込んだ本モデル（図 1）により，LSTM の中間層出力＝楽曲特徴を隠れ状態分布の形で取得でき，2 種類のデータを入力した際の分布差＝情報量が算出できるようになる．

2.2 リズム波形の特徴抽出

本稿では，グルーヴ要素として dynamics（強弱）と nuance（譜面とのタイミングずれ）の特徴を抽出対象とする．まず，元となる楽曲のドラム演奏部分のみを MIDI で作成するし，演奏音源とする．この時，dynamics と nuance を出来るだけ正確に再現する．次に，作成した MIDI 音源から dynamics と nuance を失わせた打込音源を作成する．最後に，打込音源を本モデルに学習させたのち，学習後モデルへ打込・演奏両音源を入力し隠れ状態分布の情報量差を算出する．情報量は，ある時刻 t における打込音源の分布パラメータをμ_m, σ_m，演奏音源のパラメータをμ_p, σ_pとして，Kullback-Leibler 情報量(KLD)

$$D_{KL_t}(m|p) = \log\left(\frac{\sigma_p}{\sigma_m}\right) + \frac{\sigma_m^2 + (\mu_m - \mu_p)^2}{2\sigma_p} - \frac{1}{2}$$

で算出する．

対象打音は，バス／スネア／オープンハイハットの 3 種とする．なお，演奏音源におけるグルーヴ感（ノリ感）については，dynamics は打込音源を基準としたベロシティの割合で，nuance はタイミングずれを音譜単位で表現する．

3. SOM-VAE による潜在空間の可視化

3.1 モデル概略

V. Fortuin らによる SOM-VAE は，VAE の連続的な特徴表現を離散表現にし，予め用意したトポロジカルな構造を持つ低次元空間（グリッド構造のマップ）へ写像すること

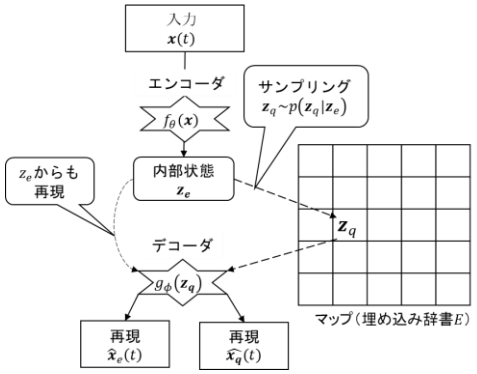

図 2　SOM-VAE モデル概略図

により，ノイズへの頑健性と視覚的な表現とを両立する手法である．モデルの概略図を図 2 に載せる．入力データ $\boldsymbol{x} \in \mathbb{R}^d$ はエンコーダ $f_\theta(\cdot)$ により潜在空間 $\boldsymbol{z_e} \in \mathbb{R}^m$ に写像される（$\boldsymbol{z_e} = f_\theta(\boldsymbol{x})$）．これはさらに，サンプリング $\boldsymbol{z_q} \sim p(\boldsymbol{z_q}|\boldsymbol{z_e})$ により，埋め込み辞書 $E = \{e_1, \ldots, e_k | e_i \in \mathbb{R}^m\}$ 内の埋め込み $\boldsymbol{z_q} \in \mathbb{R}^m$ に割り当てられる（$p(\boldsymbol{z_q}|\boldsymbol{z_e})$ はカテゴリ分布）．$\boldsymbol{z_q}$ はデコーダ $g_\phi(\cdot)$ により，$\hat{\boldsymbol{x}}_q = g_\phi(\boldsymbol{z_q})$ の形で元の入力を再現する．ここで，$\boldsymbol{z_e}$ と $\boldsymbol{z_q}$ は同じ空間にあるため，$\hat{\boldsymbol{x}}_e = g_\phi(\boldsymbol{z_e})$ の形で潜在空間からも元の入力を再現することが可能である．

3.2 時間変化の表現

SOM のマップ上での時間による変化（遷移前と遷移後のノードの近傍性）は，マルコフモデルで表される．具体的には，損失関数に遷移確率を高める損失とノード間の距離による減衰率を含めた損失の両方を含めることで時間的な近傍性を保つようにしている．

3.3 潜在空間の可視化例

MNIST の手書き文字に関し，SOM-VAE を用いて潜在空間の埋め込み $\boldsymbol{z_q}$ を可視化した例（一部）を図 3 に示す．ここでは，エンコーダ $f_\theta(\cdot)$ とデコーダ $g_\phi(\cdot)$ に 5 層の CNN（カーネルサイズ 4×4，フィルタ数 256，最大プーリング層を挟む）を用い，32×16 の SOM マップを作成している．

図では，0，2，4，6，9 の文字を認識する部分を示している．マップ上に各文字のグループが形成されていることから，SOM の自己組織化が正しく行われていることが分かる．また，4 行 4 列目のように，認識される文字と文字の境に中間状態のような画像ができていることから，時間変化についても上手く対応できていることが分かる．

以上より，潜在空間について解釈可能性の高い可視化が行えるため，グルーヴ抽出においても有用であると考えられる．

4. グルーヴ可視化に向けて

グルーヴの可視化に向けて，LSTM-VAE と SOM-VAE を組み合わせたモデルを用意する必要がある．学習に向けた順伝播・逆伝播は SOM-VAE と同様で良いと考える．エンコーダ $f_\theta(\cdot)$ には LSTM を用い，単位時間毎のリズム音源データを潜在空間への写像 $\boldsymbol{z_e}$ に変換する．$\boldsymbol{z_q} \sim p(\boldsymbol{z_q}|\boldsymbol{z_e})$ により埋め込み $\boldsymbol{z_q}$ を得，シンプルな多層 NN のデコーダ $g_\phi(\cdot)$ により $\hat{\boldsymbol{x}}_q = g_\phi(\boldsymbol{z_q})$ の形で元の入力を再現する．エンコーダ・デコーダの学習後，同じ演奏の打込音源と演奏音源をそれぞれエンコーダに入力し，マップ上での位置を可視化する事で，二つの音源の差異すなわちグルーヴの有無に関する解釈可能性の高い可視化が行われると考える．

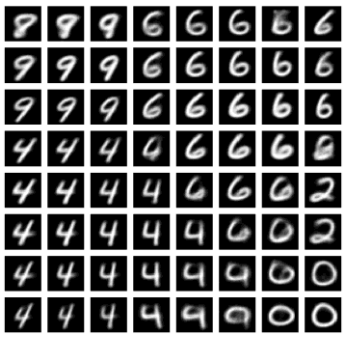

図 3　MNIST 画像のマッピング例（一部）

このモデルでは，潜在空間が離散的に表現されているので，ノイズに頑健な出力が期待できる．またこの時，打込音源と演奏音源を同時に学習させることで，グルーヴの差を視覚的に取得できるようにする．その際，打込音源と演奏音源とで明確にグループを分けるため，埋め込み辞書における初期値の分布を作為的に二分しておくことが有効であると考えられる．

5. おわりに

本研究では，VAE に SOM を組み合わせた SOM-VAE の考えを用いたノイズに頑健な特徴取得かつ可視化結果の解釈可能性の改善方法について検討した．SOM-VAE は入力をエンコーダで潜在空間へ写像した後，カテゴリ分布により SOM への離散的な埋め込みに変換し，その値を基にデコーダで元の入力を再現する．

時間的な近傍を保つようにすることで，結果として解釈可能性の高い SOM のマップが出来上がることから，グルーヴ抽出においても有用であると考えられる．打込音源と演奏音源とを同時に学習させることも可能であるが，その際は埋め込み辞書の作為的な初期化が必要であると考えられる．

参考文献

[1] 後藤真考, “技術が切り拓く音楽体験の未来”, 情報・システムソサエティ誌, vol.27, no.2, (2022).
[2] 宮丸友輔, 江村伯夫, 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, (2017).
[3] F. A. Gers, J. Schmidhuber and F. Cummins, “Learning to forget: continual prediction with LSTM”, Neural Computing. vol. 12, no. 10, (2000).
[4] D. P. Kingma, M. Welling, “Auto-Encoding Variational Bayes”, arXiv:1312.6114, (2014).
[5] S. Matsukawa, A. Suzuki, K. Arasawa and H. Matsuzaki, “Drum groove visualization using information distribution maps at LSTM variational autoencoder”, Proc. of IVSP, (2023).
[6] 松川瞬,鈴木昭弘,荒澤孔明,松﨑浩紀,“ドラムグルーヴからの振動刺激生成と印象生起に関する一検討”,情報処理学会第 87 回全国大会論文集, (2025).
[7] V. Fortuin, M. Huser, F. Locatello, H. Strathmann and G. Ratsch, “SOM-VAE: Interpretable Discrete Representation Learning on Time Series”, arXiv:1806.02199v7, (2019).