3N-01

VAEを事前分布とするNMFを用いた音楽音響信号に対するドラム譜推定

上田 舜 † 坂東 宜昭 ‡ 糸山 克寿 ‡ 吉井 和佳 ‡

† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

自動採譜は音楽情報検索技術の様々なタスクを支える基礎となっており，ドラム採譜も重要な自動採譜の1つである．一方で，ドラムパートは多様な音色・音高で表現する他の楽器パートとは異なり，スネアドラムやバスドラムなどのリズムパターンで表現する性質を持つため，一般的な楽器の自動採譜とは異なるアプローチがとられることが多い．

ドラム自動採譜においては，非負値行列因子分解（NMF）[1]や再帰型ニューラルネットワーク（RNN）[2]を用いた研究では，音色の多様性は扱われているが，リズムパターンは十分に表現されていない．一方，サポートベクターマシン（SVM）を用いて小節単位でリズムパターンを分類する研究[3]では，音楽的に妥当性の高いリズムパターンが出力されるが，それらの組み合わせなど，複雑なパターンを表現することは困難であった．したがってドラム採譜では多様な音色を表現可能な音響モデル，複雑なリズムパターンを表現可能な言語モデル，の2つが重要となる．

本稿では，楽曲の振幅スペクトログラムとそのビート時刻を入力として，ドラム譜を推定する手法について述べる．本手法は，NMFによるスペクトログラムの低ランク近似モデルをベースとして音色を表現する基底スペクトルとリズムパターンを表現する二値変数のそれぞれに変分オートエンコーダ（VAE）[4]に基づく事前分布を導入する．VAEを音響モデルの事前分布として利用することは坂東らが音声強調を目的として行って[5]おり，本研究はそれをドラム採譜に適用する．一方で，リズムパターンの事前分布，すなわち音楽における言語モデルとしてにVAEを用いることは本研究が初の試みである．

2. 提案手法

本手法はベータ過程NMF[7]を元にしており，スペクトログラム $\boldsymbol{X} \in \mathbb{R}_+^{F\times T}$ と16分音符単位のビート時刻を入力として，楽譜 $\boldsymbol{S} \in \{0,1\}^{K\times R}$，音量 $\boldsymbol{H} \in \mathbb{R}_+^{K\times T}$，音色 $\boldsymbol{W} \in \mathbb{R}_+^{F\times K}$ を推定する．$\boldsymbol{X}$ は音楽音響信号に対して調波打楽器音分離（HPSS）[6]による前処理で打楽器音のみを抽出したもの，F は周波数ビン数，T は時間フレーム数，K は基底数（楽器数），R は16分音符単位のビート数である．ビート時刻に基づき，時間フレームのインデックス $t \in \{1,\dots,T\}$ をビートのインデックス $r \in \{1,\dots,R\}$ に変換する関数 $r(t)$ を構築する．本手法の全体像を図1に示す．

ベータ過程NMFおよび上記の事前分布に基づき，スペクトログラム，楽譜，音量，音色を以下の確率的生成過程でモデル化する．

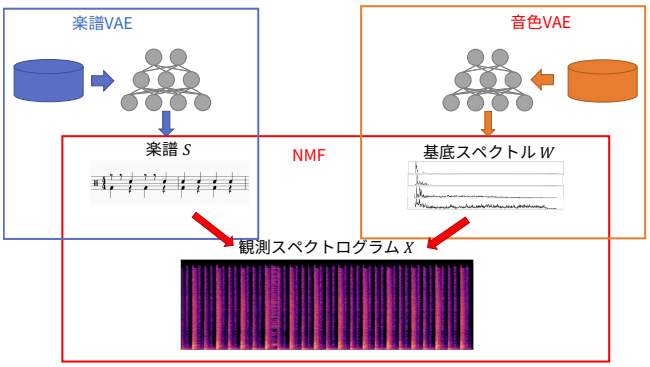

図1: モデルの全体像

$$x_{ft} \sim \mathrm{Poisson}\left(\sum_k w_{fk} h_{kt} s_{kr(t)}\right) \tag{1}$$

$$w_{fk} \sim \mathrm{Gamma}(a_{fk}, b_{fk}) \tag{2}$$

$$h_{kt} \sim \mathrm{Gamma}(c, d) \tag{3}$$

$$s_{kr} \sim \mathrm{Bernoulli}(e_{kr}) \tag{4}$$

ここで $k \in \{1,\dots,K\}$ は基底，$f \in \{1,\dots,F\}$ は周波数ビンであり，c, d は音量 $\boldsymbol{H}$ のハイパーパラメータである．以下では，楽譜 $\boldsymbol{S}$ のハイパーパラメータである e_{kr}，音色 $\boldsymbol{W}$ のハイパーパラメータである a_{fk} と b_{fk} について述べる．

2.1 楽譜 VAE

VAEを用いて楽譜 $\boldsymbol{S}$ が従う事前分布を構築する．楽譜の事前分布は1小節を単位として構築し，各小節の統計的独立性を仮定する．r 番目のビートで k 番目の楽器が発音しているか否かを表す二値変数 s_{kr} に対して以下の確率的生成モデルを考える．

$$s_{kr} \sim \mathrm{Bernoulli}(e_{kr}) \tag{5}$$

$$e_{kr} = \pi_{k,\mathrm{beat}(r)}(\boldsymbol{z}_{\mathrm{bar}(r)}) \tag{6}$$

$$\boldsymbol{z}_b \sim \mathcal{N}(0, \boldsymbol{I}) \tag{7}$$

ここで $\mathrm{bar}(r) = \lfloor \frac{r}{16} \rfloor$ はr番目のビートが何番目の小節に位置するかを表し，$\mathrm{beat}(r) = r \bmod 16$ は，r番目のビートが小節内で何番目のビートであるかを表す．$\pi_{kr'}$ は楽譜潜在変数 $\boldsymbol{z}_b$ を楽譜ベルヌーイ分布のパラメータへと変換する非線形関数であり，VAEのデコーダで構築される．

2.2 音色 VAE

VAEを用いて音色 $\boldsymbol{W}$ が従う事前分布を構築する．k 番目の楽器の f 番目の周波数インデックスの基底スペク

Estimation of Drum Scores for Music Audio Signals Based on NMF with VAE Priors: *Shun Ueda, Yoshiaki Bando, Katsutoshi Itoyama, and Kazuyoshi Yoshii (Kyoto Univ.)*

表 1: 各楽器に対する再現率，適合率，F 値

楽器名	F 値	再現率	適合率
バスドラム	0.650	0.579	0.744
スネアドラム	0.543	0.546	0.689
クローズドハイハット	0.541	0.531	0.462
オープンハイハット	0.145	0.190	0.117

トル w_{fk} に対して以下の確率的生成モデルを考える．

$$w_{fk} \sim \mathrm{Gamma}(a_{fk}, b_{fk}) \tag{8}$$
$$a_{fk} = \alpha_f(\boldsymbol{v}_k) \tag{9}$$
$$b_{fk} = \beta_f(\boldsymbol{v}_k) \tag{10}$$
$$\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{I}) \tag{11}$$

α_f と β_f は音色潜在変数 $\boldsymbol{v}_k$ をガンマ分布のパラメータへと変換する非線形関数であり，VAE のデコーダで構築される．

2.3 事後分布の推論

事後分布 $p(\mathbf{W}, \mathbf{H}, \mathbf{S}, \mathbf{Z}, \mathbf{V}|\mathbf{X})$ は解析的な計算が困難なため，ギブスサンプリングにより近似推論する．$\mathbf{W}$, $\mathbf{H}$, $\mathbf{S}$ の推論は文献 [7] と同様に行った．$\mathbf{Z}$, $\mathbf{V}$ はメトロポリスヘイスティングス法を用いて推論をする．これら 5 つの変数を交互にサンプルして事後分布を近似する．

3. 評価実験

MDB Drums データベース [8] のうち，4 曲に対して採譜し再現率，適合率および F 値によって性能評価をした．ビート情報は MDB Drums データベース内のアノテーションデータをを用い，バスドラム，スネアドラム，クローズドハイハット，オープンハイハットの 4 種類を採譜した．

楽譜 VAE の事前学習では，学習データに RWC データベース [9] のポピュラー楽曲と The Beatles，Superfly，Mr.Childlen，Aiko の曲のドラムパート譜を用いた．音色 VAE の事前学習では，RWC 楽器音データベース [10] 内のドラム楽器音を学習データに用いた．各 VAE の学習には，SGD の一種である Adam[11] を用いた．

各楽器の F 値，再現率，適合率の平均を表 1 に示す．表 1 から，楽器全体で F 値が低く，特にオープンハイハットの F 値が著しく低い結果となったことがわかる．また，正解データの一例と採譜結果を図 3 に示す．図 3 から，オープンハイハットとクローズドハイハットの区別がついていないことがわかる．また，図 2 の採譜結果が得られたときの推定した音色を図 3 に示す．図 3 から 4 つの音色全てがドラム楽器音の音色を表現できていることが分かる．また，図 2 から推定結果では正解と比べ，オンセット時刻が多く検出されている部分が存在することがわかる．これは連続したビートで楽譜が 1 となっている箇所では本来 1 音が伸びている状態であるが，ドラム音は減衰が速いため連続した 1 のそれぞれを 1 音として扱っていることが原因である．これを改善するために，バスドラム，スネアドラムとハイハットで異なる減衰モデルを設定する必要があると考える．

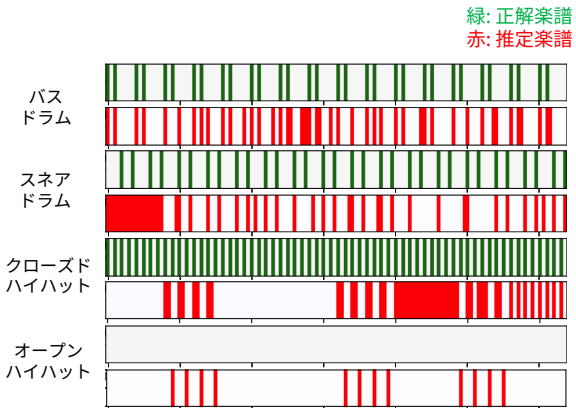

図 2: 採譜結果の例

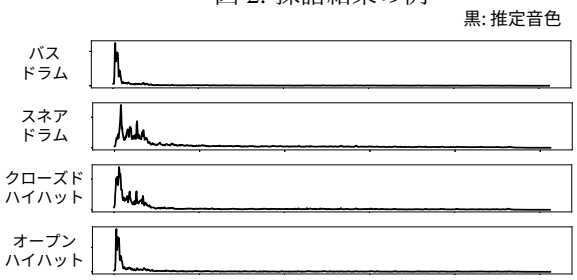

図 3: 推定された音色の例

4. おわりに

本稿では，VAE を事前分布に用いた NMF によりドラム楽譜を推定をする手法を提案した．楽譜と音色に VAE に基づく事前分布を導入することにより，より自然な分布を生成し，効率的に推論を行うことが出来る．今後は音の減衰モデルを導入しハイハット間の区別をさせることで推定結果の精度向上を図る予定である．

謝辞 本研究の一部は、JSPS 科研費 26700020, 16H01744 および JST ACCEL No. JPMJAC1602 の支援を受けた．

参考文献

[1] C.-W. Wu *et al.*: Drum Transcription Using Partially Fixed Non-Negative Matrix Factorization. *EUSIPCO*, 1281–1285, 2015.

[2] C. Southall *et al.*: Automatic Drum Transcription Using Bi-Direxional Recurrent Neural Networks. *ISMIR*, 591–597, 2016.

[3] L. Thompson *et al.*: Drum Transcription via Classification of Bar-Level Rythmic Patterns. *ISMIR*, 187–192, 2014.

[4] D. P. Kingma *et al.*: Auto-Encoding Variational Bayes. *arXiv:1312.6114*, 2013.

[5] Y. Bando *et al.*: Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization. *arXiv:1710.11439*, 2017.

[6] F. FitzGerald.: Harmonic/percussive separation using median filtering. *DAFX*, 2010.

[7] D. Liang *et al.*: Beta Process Non-negative Matrix Factorization with Stochastic Structured Mean-Field Variational Inference. *arXiv:1411.1804*, 2014.

[8] C. Southall *et al.*: MDB Drums - An Annotated Subset of MedleyDB for Automatic Drum Transcription. *ISMIR*, 2017.

[9] M. Goto *et al.*: RWC Music Database: Popular, Classical, and Jazz Music Databases. *ISMIR*, 287–288, 2002.

[10] M. Goto *et al.*: RWC Music Database: Music Genre Database and Musical Instrument Sound Database. *ISMIR*, 229–230, 2003.

[11] D. P. Kingma *et al.*: Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014.