Vol.2019-MUS-122 No.26
Vol.2019-EC-51 No.26
2019/2/23

深層ドラム譜事前分布に基づく畳み込み非負値行列因子分解を用いたドラム採譜

上田舜[1,a)]　柴田 健太郎[1,b)]　和田 雄介[1,c)]　錦見 亮[1,d)]　中村 栄太[1,e)]　吉井 和佳[1,f)]

概要：本稿では，ポピュラー音楽音響信号からあらかじめ分離されたドラム音響信号に対して，ビート時刻が既知であるという前提のもとで，バスドラム・スネアドラム・ハイハットの楽譜を推定する自動ドラム採譜について述べる．主な従来法の一つに，畳み込み非負値行列因子分解（NMFD）を用いて各時刻でのアクティベーションを推定した後，閾値処理を行うことでドラム譜を推定する方法が知られている．しかし，この手法で推定されるドラム譜は，音楽的に不自然となることを避けられなかった．この問題を解決するため，ドラム音響信号の振幅スペクトログラムに対してドラム譜の尤度を評価する NMFD に基づく音響モデルと，ドラム譜の音楽的な妥当性を評価する深層言語モデルを統合した統一的なベイズモデルを提案する．ドラム譜に対する深層言語モデルは，マルコフモデルなどの従来の確率モデルより高い表現力と汎化能力を持ち，既存のドラム譜から変分自己符号化器の枠組みで教師なし学習をすることができる．ドラム音響信号が与えられたとき，この統一的なベイズモデルとギブスサンプリングを用いてドラム譜を推定することができる．実験により，深層言語モデルの導入により採譜精度が上昇し、推定される楽譜の音楽的な自然さが改善することを確認した．

キーワード：自動採譜，ドラム採譜，言語モデル，非負値行列因子分解，変分自己符号化器

1. はじめに

ドラム自動採譜は，音楽情報検索（MIR）の分野でポピュラー音楽の採譜課題として盛んに研究されてきた [1]．採譜対象であるドラム楽器にはフロアタム，ロータム，ハイタム，ライドシンバル，クラッシュシンバルといったものが含まれているが，これまでの研究では主に，バスドラム，スネアドラム，ハイハットの 3 楽器のみが対象となっている．この理由は，これら 3 楽器がポピュラー音楽のリズムのバックボーンとして機能しているからである．音楽自動採譜の従来研究では音高をもつ楽器のオンセットとオフセットを表現するピアノロールの推定が目標となっている．これに対して，ドラム自動採譜の従来研究ではドラム楽器のオンセットを表現するドラムロールの推定が目標となっている．したがってドラム自動採譜の完成には，ドラムのオンセット時刻がドラム楽譜上のどの拍にあたるかを計算して，ドラムロールをドラム譜に変換することが必要となる．このようなプロセスは音楽自動採譜ではリズム採譜と呼ばれている [2, 3] が，ドラム自動採譜の分野においては研究されている事例はほとんど存在しない．

ドラム自動採譜では，非負値行列因子分解（NMF）がよく用いられている．これは，ドラム音の振幅スペクトログラムをドラム楽器の個数分の基底スペクトルとそれぞれの基底スペクトルに対する時間情報を持つアクティベーションの行列積に分解するものである [4–6]．NMF は音楽自動採譜によく用いられているものであり，特にドラム自動採譜には適している．この理由は，ドラム音には少数のドラム楽器による様々なドラム楽器の組み合わせパターンとそれらの音量が繰り返し含まれており，ドラム音の振幅スペクトログラムはピアノなどの音高のある楽器音の振幅スペクトログラムよりも，低ランク行列として近似しやすいからである．しかし，それぞれのドラム楽器音の音響特徴量は基底スペクトルでは完全に表現することができないので，Smaragdis [7] は NMF を畳み込み演算に拡張した畳み込み非負値行列因子分解（NMFD）を提案した．NMFD はドラム音の振幅スペクトログラムを基底スペクトログラムの重複を許したパッチワークとして近似する手法である．彼らはドラム楽器のオンセット時刻を検出するために，簡単なピークピッキングと閾値を設定し，それらを，推定したア

[1] 京都大学大学院情報学研究科
[a)] ueda@sap.ist.i.kyoto-u.ac.jp
[b)] shibata@sap.ist.i.kyoto-u.ac.jp
[c)] wada@sap.ist.i.kyoto-u.ac.jp
[d)] nishikimi@sap.ist.i.kyoto-u.ac.jp
[e)] enakamura@sap.ist.i.kyoto-u.ac.jp
[f)] yoshii@kuis.kyoto-u.ac.jp

Vol.2019-MUS-122 No.26
Vol.2019-EC-51 No.26
2019/2/23

クティベーションに適用した．このような後処理を避けるために，Liang ら [8] はベータ過程 NMF（BP-NMF）を提案した．これは二値変数（マスク）を導入することによって基底コンポーネントがその時刻に存在するかしないかを表現する手法である．

NMF やその拡張モデルは単純な採譜精度の面では有望であるが，しばしば，音楽的に不自然なドラムロールが推定される．これを回避するため，ドラムパターンの辞書をつくり，ドラム振幅スペクトログラムのセグメントをパターン辞書の登録パターンのいずれかにカテゴライズする手法が提案されている [9]．しかしこの手法は，パターン辞書に登録されていないドラムパターンを扱うことはできなかった．最近は，再帰的ニューラルネットワーク（RNN）を用いて，ドラム音のスペクトログラムから直接ドラムロールを推定するように教師あり学習を手法が提案され，推定精度が大きく向上している [10,11]．しかし，音楽的に不自然な楽譜を推定してしまうことを避けられてはいない．なぜなら，RNN ではドラム楽器の混合音の時間的なダイナミクスを学習しているが，これらは時間フレーム単位で学習されており，ビート単位で表現されるドラム譜についての考慮がされていないためである．

これらの純粋な音響モデルによる限界点を回避するため，楽譜上で定義される音楽言語モデルを導入する．このような言語モデルは近年，音楽自動採譜の分野で用いられている [12–14]．時系列による依存関係を持った音符を表現する基本的な手法として，マルコフモデルや隠れマルコフモデル（HMM）が使われている [13]．しかしこれらのモデルの表現力は限られており，また，長時間の系列依存性を扱うと，マルコフモデルは計算が困難となる問題も存在した．RNN に基づく言語モデルは，近年，楽譜の長時間にまたがる依存関係を学習するために，ピアノロールから楽譜を推定する際に，NMF のような低ランク近似による音響モデルとともに用いられていた [14]．離散シンボルで定義された言語モデルと連続な変数で定義された音響モデルを原則的に統合することは，未解決の問題となっている．

本稿では，ビート時刻（16 分音符単位）と小節が始まる時刻が予め与えられていると仮定した上で（これらはビートトラッキング手法により推定できる [15]），DNN に基づく言語モデルと NMFD に基づく音響モデルを統合したベイズモデル（図 1）に基づいたドラム自動採譜の新たな手法を提案する．音響モデルではドラム音のスペクトログラムに対するドラム譜（ビート単位の二値変数）の尤度を評価し，言語モデルではドラム譜の事前確率（音楽的妥当性）を評価する．ドラム音の音響的加法性は NMFD に基づく線形モデルによって表現できるが，ドラム譜の文法的構造は複雑であり，緻密に表現することは難しい．したがって，我々はドラムパターンデータから教師なし学習の手法である変分自己符号化器（VAE）[16] の枠組みを用いて，1 小節のドラムパターンの生成モデルと潜在的な特徴表現を学習した．観測としてドラム音のスペクトログラムが与えられると，ドラム譜 1 小節のドラムパターンが連続したものと音響モデルと言語モデルのすべての変数がギブスサンプリングに基づいて推定できる．

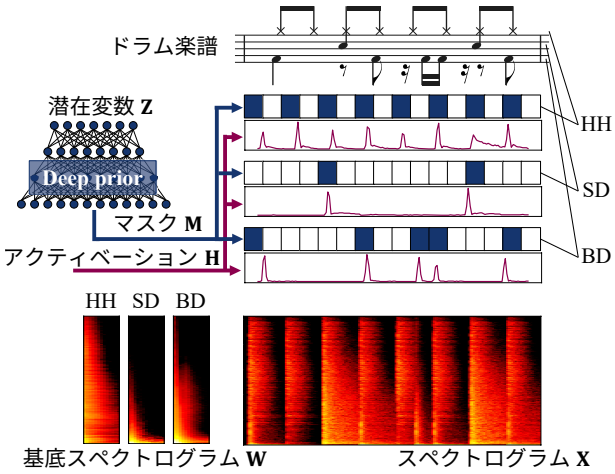

図 1　DNN に基づくドラム譜モデルと NMFD に基づく音響モデルを統合したドラムパート振幅スペクトログラムの階層的生成モデル．HH，SD，BD はそれぞれハイハット，スネアドラム，バスドラムを表す．

本研究の重要な部分は，ベイズ学習の統計的枠組みと深層学習の表現力との統合である．本研究は，ドラム自動採譜に強力な深層生成モデルを用いる初めての試みであり，この手法は一般的な音楽自動採譜にも応用できる．提案する深層ベイズアプローチの主な利点は，End-to-End 学習で必要であった音響信号と正解データのペアデータを必要とせず，既存楽曲で用いられている大量のドラム譜データ [17] によって言語モデルが学習できることである．

2. 提案法

本章では，ポピュラー音楽の音響信号から抽出したドラムパート音響信号からドラム譜を推定する手法を述べる．

2.1 問題設定

本稿で扱うドラム自動採譜問題を以下で定義する．

入力: ドラムパート音響信号の振幅スペクトログラム $\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ と 16 分音符単位のビート時間と小節線情報
出力: ドラム譜　$\mathbf{S} \in \{0,1\}^{K \times R}$.

ここで，F は周波数ビン数，T は時間フレーム数，$K = 3$ はドラムの楽器数（バスドラム，スネアドラム，ハイハット），R は観測信号におけるビート数である．入力となる音響信号は HPSS [18] によってあらかじめ分離されており，打楽器音のみで構成されると仮定する．二値変数 S_{kr} は，ドラム k が r 番目のビートでオンセットを持つかどうかを表し，小節単位のドラムパターンに分割できる．

2.2 モデルの定式化

DNN に基づく $\mathbf{S}$ の言語モデルと NMFD に基づく $\mathbf{X}$ の音響モデルを統合し，振幅スペクトログラム $\mathbf{X}$ の階層的生成モデルを定式化する（図 1）．

2.2.1 NMFD に基づく音響モデル（楽譜尤度）

振幅スペクトログラム $\mathbf{X}$ は，基底スペクトログラム $\mathbf{W} \in \mathbb{R}_{+}^{(K+1)\times F\times M}$，アクティベーションベクトル $\mathbf{H} \in \mathbb{R}_{+}^{(K+1)\times T}$，二値マスク $\mathbf{S} \in \{0,1\}^{K\times R}$ を用いて次式で近似できる．

$$X_{ft} \approx Y_{ft} \stackrel{\text{def}}{=} \sum_{m=1}^{M}\sum_{k=0}^{K} Y_{ftkm} \tag{1}$$

ここで，Y_{ftkm} は次式で与えられる．

$$\begin{cases} Y_{ftkm} = W_{kfm}H_{k,t-m}S_{k,r(t-m)} & (k \geq 1) \\ Y_{ft0m} = W_{0fm}H_{0,t-m} \end{cases} \tag{2}$$

ここで，M はそれぞれの基底スペクトログラムの持つ時間フレーム数であり，$\{W_{kfm}\}_{f=1}^{F}$ $(k \geq 1)$ はドラム k のフレーム m における基底スペクトルである．また，$r(t)$ は時間フレーム t が所属するビートの位置を表す．追加の基底スペクトログラム W_{0fm} とそのアクティベーションベクトル H_{0t} を導入する．これらは入力されたドラム音に含まれる雑音成分を表現する．式 (1) の近似誤差を評価するために，KL-NMF [19] と同様に，Kullback-Leibler（KL）ダイバージェンスを用いる．確率的な見地に立てば，KL ダイバージェンスの最小化は式 (3) のポアソン尤度の最大化と等価である．

$$X_{ft} \sim \text{Poisson}(Y_{ft}) \tag{3}$$

ベイズモデルとしてで定式化するため，以下のように $\mathbf{W}$ の事前分布にポアソン尤度の共役事前分布であるガンマ分布を用いる．

$$\begin{cases} W_{kfm} \sim \text{Gamma}(a_{kfm}, b_{kfm}) & (k \geq 1) \\ W_{0fm} \sim \text{Gamma}(a_0, b_0) \end{cases} \tag{4}$$

ここで $\text{Gamma}(a_*, b_*)$ は形状母数 a_* と逆尺度母数 b_* を持つガンマ分布を表し，a_* と b_* はハイパーパラメータである．同様に，$\mathbf{H}$ の事前分布にもガンマ分布を用いる．

$$\begin{cases} H_{kt} \sim \text{Gamma}(c_k, d_k) & (k \geq 1) \\ H_{0t} \sim \text{Gamma}(c_0, d_0) \end{cases} \tag{5}$$

ここで c_k，d_k，c_0，d_0 はハイパーパラメータである．

2.2.2 DNN に基づく言語モデル（楽譜事前分布）

二値マスク $\mathbf{S}$ は以下のベルヌーイ分布によって各要素が独立に生成されると仮定する．

$$S_{kr} \sim \text{Bernoulli}(\pi_{kr}) \tag{6}$$

ここで，π_{kr} はドラム k が r 番目のビートにオンセットを持つ確率である．式 (6) はドラムと小節単位で書き直し，次式で表現できる．

$$\mathbf{s}_i \sim \text{Bernoulli}(\boldsymbol{\pi}_i) \tag{7}$$

ここで $\mathbf{s}_i$ と $\boldsymbol{\pi}_i$ はそれぞれ，$16K$ 次元の二値ベクトルと実数ベクトルであり，小節 i $(0 \leq i \leq I-1)$ に含まれる S_{kr} と π_{kr} で構成される．提案法の核は，$\boldsymbol{\pi}_i$ を深層生成モデルで表現することにある．

$$\mathbf{z}_i \sim \mathcal{N}(0,1) \tag{8}$$

$$\boldsymbol{\pi}_i = \text{DNN}_\theta(\mathbf{z}_i) \tag{9}$$

ここで，DNN_θ は θ をパラメータにもつ非線形関数であり，$\mathbf{z}_i$ を $\boldsymbol{\pi}_i$ に変換する．$\mathbf{z}_i$ は小節 i での V 次元の潜在変数である．簡単のため，$\text{bar}(r) = \lfloor \frac{r}{16} \rfloor$ をビート r が所属する小節を表す関数とし，$\text{tatum}(r) = r \bmod 16$ をビート r が小節内のどの位置のビートであるかを表す関数とする．楽譜の事前分布 $p_\theta(\mathbf{S})$ は $p_\theta(\mathbf{S}|\mathbf{Z})p(\mathbf{Z})$ によって与えられる生成モデルを潜在変数 $\mathbf{Z}$ で周辺化することで得られる．

2.3 楽譜事前分布の学習

深層楽譜事前分布 $p_\theta(\mathbf{S})$ を計算するため，既存ドラムパターン $\mathbf{S}$ を用いて変分自己符号化器（VAE）を教師なし学習する．我々の目標は，DNN パラメータ θ を $p_\theta(\mathbf{S})$ で表される尤度を最大化するように推定することである．しかし $p_\theta(\mathbf{S})$ の直接的な最大化は困難であるため，$p_\theta(\mathbf{S})$ の下限を最大化することで $\log p_\theta(\mathbf{S})$ を間接的に最大化する．具体的には，任意の分布 $q(\mathbf{Z})$ を導入し，イエンセンの不等式を用いることで，変分下限を導出する．

$$\log p_\theta(\mathbf{S}) \geq -\mathbb{KL}[q(\mathbf{Z})|p(\mathbf{Z})] + \mathbb{E}_q[\log p_\theta(\mathbf{S}|\mathbf{Z})] \tag{10}$$

$q(\mathbf{Z})$ の代わりに，認識モデル $q_\phi(\mathbf{Z}|\mathbf{S})$ をパラメータ ϕ を用いて次式で定義する．

$$q_\phi(\mathbf{Z}|\mathbf{S}) = \prod_{i=0}^{I-1} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_\phi(\mathbf{s}_i), \boldsymbol{\sigma}_\phi^2(\mathbf{s}_i)) \tag{11}$$

ここで $\boldsymbol{\mu}_\phi$ と $\boldsymbol{\sigma}_\phi^2$ はそれぞれ，入力と出力が V 次元と $16K$ 次元の DNN によって定義される非線形関数である．さらに $\log p_\theta(\mathbf{S})$ の下限は次式で書くことができる．

$$\begin{aligned} &\log p_\theta(\mathbf{S}) \\ &\geq \frac{1}{2}\sum_{i,v}\left(1 + \log \boldsymbol{\sigma}_{\phi,v}^2(\mathbf{s}_i) - \boldsymbol{\mu}_{\phi,v}^2(\mathbf{s}_i) - \boldsymbol{\sigma}_{\phi,v}^2(\mathbf{s}_i)\right) \\ &\quad + \sum_{k,r}\mathbb{E}_q[S_{kr}\log\pi_{kr} + (1-S_{kr})\log(1-\pi_{kr})] \end{aligned} \tag{12}$$

ここで $\boldsymbol{\mu}_{\phi,v}^2(\mathbf{s}_i)$ は $\boldsymbol{\mu}_\phi^2(\mathbf{s}_i)$ の v 番目の次元である．$\boldsymbol{\sigma}_{\phi,v}^2(\mathbf{s}_i)$ も同様である．式 (12) は θ と ϕ の関数である．なぜなら $\boldsymbol{\pi}$ は式 (9) で既に定義されているからである．θ と ϕ はともに式 (12) で与えられる下限を Adam [20] などの確率的勾配降下法で最大化することにより同時に最適化される．

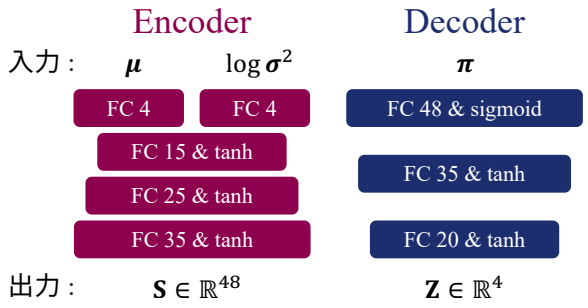

図 2　ドラムパターンに対する VAE.

2.4　事後分布の計算

観測データとして $\mathbf{X}$ が与えられたとき，我々の目標は事後分布 $p(\mathbf{W}, \mathbf{H}, \mathbf{S}, \mathbf{Z}|\mathbf{X})$ を計算することである．しかしこの事後分布は解析的な計算が困難なため，ギブスサンプリングを用いて $\mathbf{W}$，$\mathbf{H}$，$\mathbf{S}$，$\mathbf{Z}$ を交互にサンプリングする．

2.4.1　ドラム譜の更新

音響モデルに含まれる $\mathbf{W}$ と $\mathbf{H}$ と言語モデルに含まれる $\mathbf{Z}$ を用いて，二値変数 $\mathbf{S}$ をサンプルする．

$$S_{kr} \sim \text{Bernoulli}\left(\frac{\mathrm{P}^1_{kr}}{\mathrm{P}^0_{kr} + \mathrm{P}^1_{kr}}\right) \tag{13}$$

$$\mathrm{P}^0_{kr} \propto (1 - \pi_{kr})\, p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 0) \tag{14}$$

$$\mathrm{P}^1_{kr} \propto \pi_{kr}\, p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 1) \tag{15}$$

ここで，式 (14) や式 (15) の第一項と第二項はそれぞれ言語モデルの事前確率と音響尤度を表す．$\mathbf{S}_{\neg(kr)}$ は $\mathbf{S}$ から S_{kr} を除いた部分集合を表し，$\boldsymbol{\pi}$ は $\mathbf{Z}$ に依存する．式 (14) と式 (15) の尤度項は次式で与えられる．

$$\begin{aligned} &p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 0) \\ &= \prod_{t \in \{r(t)=r\}} \prod_{f} \left(Y^{\neg k}_{ft} + \sum_m W_{kfm} H_{k,t-m}\right)^{X_{ft}} \\ &\qquad \cdot \exp\left\{-\sum_m W_{kfm} H_{k,t-m}\right\} \end{aligned} \tag{16}$$

$$\begin{aligned} &p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 1) \\ &= \prod_{t \in \{r(t)=r\}} \prod_{f} \left(Y^{\neg k}_{ft}\right)^{X_{ft}} \end{aligned} \tag{17}$$

ここで，$Y^{\neg k}_{ft}$ は次式で与えられる．

$$Y^{\neg k}_{ft} = \sum_{l \neq k} \sum_m Y_{ftkm} \quad (k \geq 1) \tag{18}$$

2.4.2　NMFD に基づく音響モデルの更新

二値変数 $\mathbf{S}$ を含むベイズ NMFD に含まれる $\mathbf{W}$，$\mathbf{H}$，$\mathbf{S}$ のサンプリングは，BP-NMF [8] と呼ばれる二値変数ベイズ NMF のギブスサンプリング手法を拡張した手法によりサンプリングする．具体的には，$\mathbf{H}$ と $\mathbf{S}$ が与えられた状態での $\mathbf{W}$ のサンプリングは次式で与えられる．

$$W_{kfm} \sim \text{Gamma}(\hat{a}_{kfm}, \hat{b}_{kfm}) \tag{19}$$

$$\begin{cases} \hat{a}_{kfm} = \sum_t X_{ft} \lambda_{ftkm} + a_{kfm} & (k \geq 1) \\ \hat{a}_{0fm} = \sum_t X_{ft} \lambda_{ft0m} + a_0 \end{cases} \tag{20}$$

$$\begin{cases} \hat{b}_{kfm} = \sum_t H_{k,t-m} S_{k,t-m} + b_{kfm} & (k \geq 1) \\ \hat{b}_{0fm} = \sum_t H_{0,t-m} + b_0 \end{cases} \tag{21}$$

ここで，λ_{ftkm} は補助変数であり，次式で与えられる．

$$\lambda_{ftkm} = \frac{Y_{ftkm}}{Y_{ft}} \tag{22}$$

同様に，$\mathbf{W}$ と $\mathbf{S}$ が与えられた状態で，$\mathbf{H}$ のサンプリングを行う．

$$H_{kt} \sim \text{Gamma}(\hat{c}_{kt}, \hat{d}_{kt}) \tag{23}$$

$$\begin{cases} \hat{c}_{kt} = \sum_{f,m} X_{ft} \lambda_{f,t+m,km} + c_k & (k \geq 1) \\ \hat{c}_{0t} = \sum_{f,m} X_{ft} \lambda_{f,t+m,0m} + c_0 \end{cases} \tag{24}$$

$$\begin{cases} \hat{d}_{kt} = \sum_{f,m} W_{kfm} S_{kt} + d_k & (k \geq 1) \\ \hat{d}_{0t} = \sum_{f,m} W_{0fm} + d_0 \end{cases} \tag{25}$$

2.4.3　DNN に基づく言語モデルの更新

潜在変数 $\mathbf{Z}$ の事後分布の解析的な導出は困難であるため，メトロポリス・ヘイスティングス法によって $\mathbf{Z}$ を更新する．提案変数 $\mathbf{z}^*_i$ はそれぞれの小節 i ごとに次式のガウス分布によるランダムウォークによって生成される．

$$\mathbf{z}^*_i \sim q(\mathbf{z}^*_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i, 0.1) \tag{26}$$

提案変数 $\mathbf{z}^*_i$ は次式の受理率 $a_{\mathbf{z}^*_i|\mathbf{z}_i}$ によって次回反復時の $\mathbf{z}_i$ として受理される．

$$a_{\mathbf{z}^*_i|\mathbf{z}_i} = \min\left(1, \frac{p(\mathbf{z}^*_i)}{p(\mathbf{z}_i)} \prod_{k,r \in \{\text{bar}(r)=i\}} \frac{p(S_{kr}|\mathbf{z}^*_i)}{p(S_{kr}|\mathbf{z}_i)}\right) \tag{27}$$

適切な $\mathbf{Z}$ を推定するため，メトロポリス・ヘイスティングス法の初期値として，推定した $\mathbf{S}$ と認識モデル $q_\phi(\mathbf{Z}|\mathbf{S})$ からサンプリングした $\mathbf{Z}$ を用いる．

3.　評価実験

3.1　実験設定

実験には，RWC ポピュラー音楽データベース [21] の音響信号を使用した．これらの音響信号をモノラル信号に変換した後に，30 秒ごとのセグメントに分割した．評価には開始 30 秒から開始 60 秒までの 30 秒間を用いた．データベースの 100 曲の中から，評価に用いた 30 秒間のなかにバスドラム，スネアドラム，ハイハットがそれぞれ少なくとも 1 回演奏されている 64 曲を使用した．16 分音符単位のビート時間情報と小節線情報はアノテーションデータ [22] を使用した．ビート時刻はアノテーションとして与えられたものを 0.03 秒早い方向にシフトさせ，ドラム音の実際のオンセット時刻とビート時刻とのわずかなずれを吸収した．

RWC ポピュラー音楽データベース内のすべての曲は 44.1kHz でサンプリングされている．これらの音響信号から短時間フーリエ変換（STFT）を用いて複素スペクトログラムを計算した．この際，窓関数はハン窓を用い，窓幅は 2048 サンプル，シフト長は 441 サンプル（10 ミリ秒）とした．その後，HPSS [18] を用いてドラム音の振幅スペクトログラムを分離した．各曲それぞれの振幅スペクトログラムは，全要素の平均が 1 となるように正規化した．

ハイパーパラメータ a_{kfm}，b_{kfm} $(k \geq 1)$ を決定するために，RWC 楽器音データベース [23] からバスドラム，スネアドラム，ハイハットそれぞれの楽器音を用いた．それぞれの楽器音振幅スペクトログラムには上記の通り類似性があり，各ドラム楽器においてデータベース内の該当楽器音を連結させた振幅スペクトログラムを作成し，1 基底による NMFD を適用させて得た基底スペクトログラムを，その楽器のテンプレートスペクトログラムとした．このようにして得られたテンプレートスペクトログラムを平均として，分散を 0.01 となるようなガンマ分布のパラメータをハイパーパラメータ a_{kfm}，b_{kfm} $(k \geq 1)$ として決定した．他の $\mathbf{W}$ とアクティベーションベクトル $\mathbf{H}$ の事前分布のハイパーパラメータは以下のように設定した．$a_{0fm} = 0.05$，$b_{0fm} = 50.0$，$c_0 = 50.0$，$d_0 = 50.0$，$c_k = 1.0$，$d_k = 50.0$．VAE ネットワークの学習には，ビートルズや日本のポピュラー音楽のドラム譜から得た 41474 小節を使用した．これらのデータはテストデータと重複していない．基底スペクトログラムのフレーム数は $M = 20$ とし，潜在変数 $\mathbf{z}_i$ の次元数は $V = 4$ とした．

ドラム自動採譜の性能は以下で定義される適合率，再現率，F 値によって測定した．

$$\mathcal{P} = \frac{N_c}{N_e}, \quad \mathcal{R} = \frac{N_c}{N_g}, \quad \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}} \tag{28}$$

ここで N_e，N_g，N_c はそれぞれ推定結果の音符数，正解データの音符数，正解した音符数である．それぞれの楽器 k (≥ 1) について，オンセット t^* を，推定した変数 $\mathbf{H}$ と $\mathbf{S}$ から以下の条件に沿うように検出した．

$$H_{kt^*}S_{kt^*} \geq 0.3 \cdot \max_t\{H_{kt}S_{kt}\} \tag{29}$$

$$H_{kt^*}S_{kt^*} = \max_{t^*-5\leq t\leq t^*+5}\{H_{kt}S_{kt}\} \tag{30}$$

このとき，推定したオンセットと正解データのオンセットの時刻差が 50 ミリ秒以内ならば，正解であるとした．

3.2 実験結果

実験結果を表 1 に示す．スネアドラムとハイハットに関して，提案法はすべての指標で NMFD を上回った．バスドラムに関しては，再現率が NMFD よりわずかに悪化したものの，F 値は同程度となった．採譜例を図 3 に示す，スネアドラムに関して，NMFD（音響モデル）のみで推定した結果には，音楽的に不自然なリズムパターンが含まれていた（たとえば最後の半小節）．一方，提案法では自然な採譜結果が得られた．このことから，DNN に基づく言語モデルと NMFD に基づく音響モデルを統合することで，採譜精度が改善するだけでなく，音楽的な自然さも向上することが確認できた．

手法	パート	$\mathcal{P}$(%)	$\mathcal{R}$(%)	$\mathcal{F}$(%)
NMFD	HH	79.4	60.9	69.0
	SD	63.2	63.6	63.4
	BD	82.3	80.2	81.2
VAE-NMFD	HH	80.9	61.4	69.8
	SD	67.6	65.4	66.5
	BD	83.0	79.4	81.2

表 1 RWC ポピュラー音楽データベースでのドラム自動採譜の結果．HH，SD，BD はそれぞれハイハット，スネアドラム，バスドラムを表す．

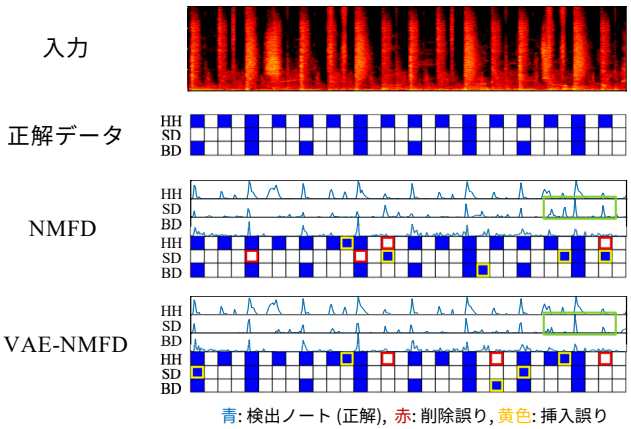

図 3 NMFD と VAE-NMFD のそれぞれによって推定されたドラム譜の例．VAE-NMFD ではアクティベーションはマスクを適用したあとのものである．

4. おわりに

本稿では，NMFD に基づく音響モデルと VAE に基づく深層生成モデルを確率的に統合したドラムの自動採譜手法を提案した．本手法は，言語モデルを楽譜データのみから学習することができ，DNN に基づく標準的な End-to-End 学習で必要とされる時系列のペアデータを必要としない．実験では，楽譜の深層生成モデルを用いることで，音楽的に自然な楽譜が推定されることが確認できた．今後は，[24] のように，提案法とビート時刻・小節線時刻の検出手法を組み合わせ，ドラム譜とビート時刻を同時に推定することに取り組みたい．またドラムパターンの時系列の依存関係や繰り返し構造を表現し，VAE を時系列で再帰的に扱うことも有効であると考えられる．

謝辞 本研究の一部は JST ACCEL No. JPMJAC1602，科研費 No. 26700020，16H01744，16J05486 の支援を受けた．

参考文献

[1] C. Wu, C. Dittmar, C. Southall and R. Vogl: A Review of Automatic Drum Transcription, *Journal of IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 9, pp. 1457–1483 (2018).

[2] C. Raphael: A Hybrid Graphical Model for Rhythmic Parsing, *Artificial Intelligence*, Vol. 137, pp. 217–238 (2002).

[3] E. Nakamura, K. Yoshii and S. Sagayama: Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 4, pp. 794–806 (2017).

[4] J. Paulus and T. Virtanen: Drum transcription with non-negative spectrogram factorisation, *13th European Signal Processing Conference, (EUSIPCO)*, pp. 1–4 (2005).

[5] C. Dittmar and D. Gärtner: Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition, *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, pp. 187–194 (2014).

[6] C. Wu and A. Lerch: Drum Transcription Using Partially Fixed Non-Negative Matrix Factorization with Template Adaptation, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 257–263 (2015).

[7] P. Smaragdis: Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs, *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, pp. 494–499 (2004).

[8] D. Liang, M. D. Hoffman and D. P. W. Ellis: Beta process sparse nonnegative matrix factorization for music, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 375–380 (2013).

[9] L. Thompson, S. Dixon and M. Mauch: Drum Transcription via Classification of Bar-Level Rhythmic Patterns, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 187–192 (2014).

[10] C. Southall, R. Stables and J. Hockman: Automatic Drum Transcription Using Bi-Directional Recurrent Neural Networks, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 591–597 (2016).

[11] R. Vogl, M. Dorfer and P. Knees: Drum transcription from polyphonic music with recurrent neural networks, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 201–205 (2017).

[12] S. Sigtia, E. Benetos and S. Dixon: An end-to-end neural network for polyphonic piano music transcription, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 24, No. 5, pp. 927–939 (2016).

[13] R. Nishikimi, E. Nakamura, M. Goto, K. Itoyama and K. Yoshii: Scale- and Rhythm-Aware Musical Note Estimation for Vocal F0 Trajectories Based on a Semi-Tatum-Synchronous Hierarchical Hidden Semi-Markov Model, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 376–382 (2017).

[14] A. Ycart and E. Benetos: Polyphonic Music Sequence Transduction with Meter-Constrained LSTM Networks, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018).

[15] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs and G. Widmer: Madmom: A new Python audio and music signal processing library, *Proceedings of the ACM International Conference on Multimedia*, pp. 1174–1178 (2016).

[16] D. P. Kingma and M. Welling: Auto-encoding variational Bayes, *Proceedings of the International Conference on Learning Representations (ICLR)* (2014).

[17] C. Raffel: "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching", PhD Thesis, COLUMBIA UNIVERSITY (2016).

[18] D. Fitzgerald: Harmonic/Percussive Separation Using Median Filtering, *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, pp. 1–4 (2010).

[19] P. Smaragdis and J. C. Brown: Non-negative matrix factorization for polyphonic music transcription, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 177–180 (2003).

[20] D. P. Kingma and J. Ba: Adam: A Method for Stochastic Optimization, *arXiv*, Vol. 1412.6980 (online), available from ⟨http://arxiv.org/abs/1412.6980⟩ (2014).

[21] M. Goto, H. Hashiguchi, H. Nishimura and R. Oka: RWC Music Database: Popular, Classical and Jazz Music Databases, *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pp. 287–288 (2002).

[22] M. Goto: AIST Annotation for the RWC Music Database., *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pp. 359–360 (2006).

[23] M. Goto, H. Hashiguchi, H. Nishimura and R. Oka: RWC Music Database: Music genre database and musical instrument sound database, *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pp. 229–230 (2003).

[24] R. Vogl, M. Dorfer, G. Widmer and P. Knees: Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 150–157 (2017).