5S-05

# 深層音響・言語モデルの統合に基づくドラム採譜

石塚 崚斗　上田 舜　錦見 亮　中村 栄太　吉井 和佳
京都大学 大学院情報学研究科

## 1. はじめに

ドラム採譜は，音楽音響信号からドラムを構成する各楽器の発音時刻を検出するタスクである．従来，非負値行列因子分解 (NMF) [1] がしばしば用いられていたが，解析対象に合致する基底スペクトルを準備・学習することは容易ではなかった．最近，再帰型ニューラルネットワーク (RNN) [2] の利用により，推定精度は向上しつつある．しかし，推定結果に音楽的に不自然な箇所が含まれる問題が依然としてある．

ドラム採譜では，ドラム譜の音響信号に対する当てはまりの良さを表す音響モデルと，ドラム譜の音楽的な妥当性を評価する言語モデルとを統合する手法が有効であると考えられる．上田ら [3] は，畳込み NMF に基づく音響モデルと変分オートエンコーダ (VAE) に基づく深層言語モデルの統合に基づくドラム採譜手法を提案している．本稿では，畳込みニューラルネットワーク (CNN) に基づく深層音響モデル [4] と VAE に基づく深層言語モデルを統合するドラム採譜手法を提案する．これまで，本研究のように，深層音響モデルと深層言語モデルを統合する試みは存在しなかった．

## 2. 提案法

時間-周波数分解能が異なる $C$ 個の振幅スペクトログラム (MCMS) $\mathbf{X} \in \mathbb{R}^{C\times F\times T}$ と，テイタム時刻系列 $\{b_m\}_{m=1}^{M}$（16 分音符単位）を入力として，ドラム譜を推定する（図 1）．ここで，$C$ はチャネル数，$F$ は対数スケール周波数ビン数，$T$ は時間フレーム数，$M$ は入力楽曲のテイタム数を表す．CNN に基づく音響モデルと VAE に基づく言語モデルはそれぞれ独立に事前学習しておき，両者の評価値の和を最大化するよう，テイタム単位のドラム譜 $\mathbf{Y} \in \{0,1\}^{K\times M}$ を最適化することで最終出力のドラム譜を得る（$K$ はドラムの種類数）．

### 2.1 音響モデルの学習

CNN は，振幅スペクトログラム $\mathbf{X} \in \mathbb{R}^{C\times F\times T}$ を入力として，フレーム単位のアクティベーション $\mathbf{H} \in [0,1]^{K\times T}$ を出力する．$H_{kt}$ は $t$ 番目のフレームにドラム $k$ の発音がある確からしさを表す．学習時には，正解の発音時刻ラベルを $\hat{\mathbf{Y}} \in \{0,1\}^{K\times T}$ とすると，以下のベルヌイ確率を最大化 (クロスエントロピーの最小化に相当) するよう CNN のパラメータを推定する．

$$\mathcal{J}_{\mathrm{CNN}} = \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}\big(a\hat{Y}_{kt}\log H_{kt} + (1-\hat{Y}_{kt})\log(1-H_{kt})\big) \quad (1)$$

ここで，$a \in \mathbb{R}_+$ は正例に対する重みである．

採譜を行う際には，学習済みの CNN を用いて入力スペクトログラム $\mathbf{X}$ からフレーム単位のアクティベーション $\mathbf{H}$ を推定し，各テイタム区間における最大値を取ることでテイタム単位のアクティベーション $\mathbf{G} \in [0,1]^{K\times M}$ に変換する．最適化対象となるドラム譜 $\mathbf{Y} \in \{0,1\}^{K\times M}$ に対して，音響的な当てはまりの良さを評価する音響モデルを，以下のベルヌイ確率に基づく評価関数で定める．

$$\mathcal{J}_{\mathrm{acoust}}(\mathbf{Y}) = \frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M}\big(Y_{km}\log G_{km} + (1-Y_{km})\log(1-G_{km})\big) \quad (2)$$

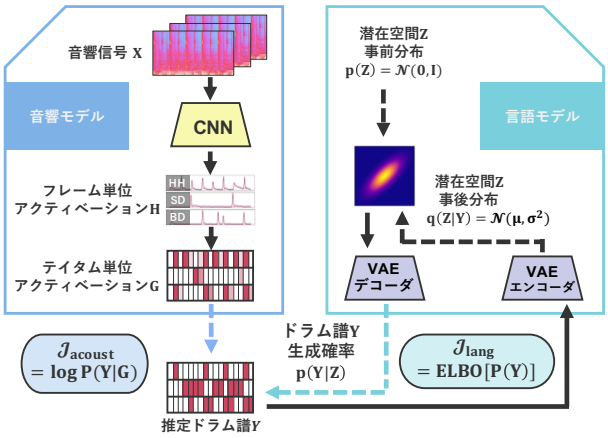


図 1: 深層音響・言語モデルに基づくドラム採譜

### 2.2 言語モデルの学習

VAE は潜在変数モデルであり，小節単位のドラム譜の生成モデルとして用いる．学習時には，テイタム単位のドラム譜 $\mathbf{B} \in \{0,1\}^{K\times M}$ の学習データに対して，周辺尤度 $p(\mathbf{B})$ を最大化するようにパラメータを推定する．実際には，潜在変数 $\mathbf{Z} \in \mathbb{R}^{D}$（$D$ は潜在変数の次元）に対する変分事後分布 $q(\mathbf{Z}|\mathbf{B})$ を導入して得られる次の変分下限を最大化することで，間接的に $p(\mathbf{B})$ を最大化する．

$$\begin{aligned}\log p(\mathbf{B}) &\geq \int q(\mathbf{Z}|\mathbf{B})\log\frac{p(\mathbf{B},\mathbf{Z})}{q(\mathbf{Z}|\mathbf{B})}d\mathbf{Z}\\ &= \mathbb{E}_q[\log p(\mathbf{B}|\mathbf{Z})] - \mathrm{KL}[q(\mathbf{Z}|\mathbf{B})|p(\mathbf{Z})]\\ &= \mathcal{J}_{\mathrm{lang}}(\mathbf{B}) \end{aligned} \quad (3)$$

ただし，KL[·] は KL ダイバージェンスを表し，$p(\mathbf{B}|\mathbf{Z})$，$p(\mathbf{Z})$，$q(\mathbf{Z}|\mathbf{B})$ には以下の分布を仮定する．

$$p(\mathbf{B}|\mathbf{Z}) = \mathrm{Bernoulli}(\boldsymbol{\pi}(\mathbf{Z})) \quad (4)$$

$$p(\mathbf{Z}) = \mathcal{N}(\mathbf{Z};\mathbf{0},\mathbf{I}) \quad (5)$$

$$q(\mathbf{Z}|\mathbf{B}) = \mathcal{N}(\mathbf{Z};\boldsymbol{\mu}(\mathbf{B}),\boldsymbol{\sigma}^2(\mathbf{B})) \quad (6)$$

$\boldsymbol{\pi}(\mathbf{Z}) \in [0,1]$ はデコーダの出力，$\mathbf{I} \in \mathbb{R}^{D\times D}$ は単位行列，$\boldsymbol{\mu}(\mathbf{B}) \in \mathbb{R}^{D}$，$\boldsymbol{\sigma}^2(\mathbf{B}) \in \mathbb{R}_+^{D\times D}$ はエンコーダの出力を表す．

採譜を行う際には，最適化対象となるドラム譜 $\mathbf{Y}$ に対して音楽的な妥当性を評価するため，$\mathbf{Y}$ の対数周辺尤

---

Automatic Drum Transcription Based on Integration of Deep Acoustic and Language Models: *Ryoto Ishizuka, Shun Ueda, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii (Kyoto Univ.)*



Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.

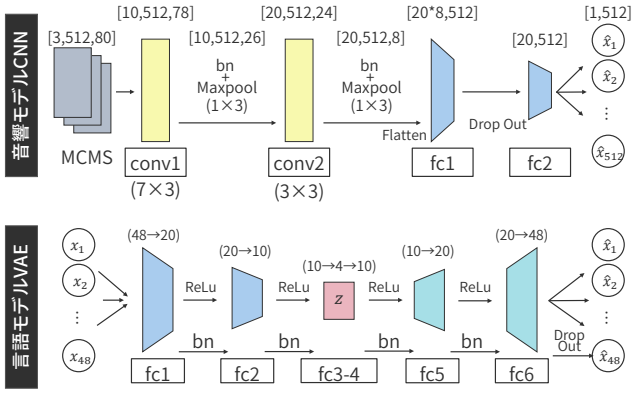


図 2: CNN と VAE のネットワーク構造

度の変分下限 $\mathcal{J}_{\mathrm{lang}}(\mathbf{Y})$ を評価関数として用いる．

### 2.3 音響・言語モデルに基づくドラム譜の最適化

最大化すべき評価関数は，音響モデルと言語モデルの重み付き和で定める．

$$\mathcal{J}_{\mathrm{total}}(\mathbf{Y}) = \mathcal{J}_{\mathrm{acoust}}(\mathbf{Y}) + \alpha \mathcal{J}_{\mathrm{lang}}(\mathbf{Y}) \tag{7}$$

ここで，$\alpha \in \mathbb{R}_+$ は言語モデルの重みである．$\mathbf{Y}$ はバイナリデータであるので，これは組合せ最適化問題になっており，通常の勾配に基づく反復最適化はできない．本研究では，最適化対象のドラム譜 $\mathbf{Y} \in \{0,1\}^{K\times M}$ を，学習済み CNN の出力 $\mathbf{G}$ に対して閾値処理を行ったもので初期化したあと，各要素 $Y_{km}$ をランダムな順序でフリップし，$\mathcal{J}_{\mathrm{total}}$ を逐次最大化する．

## 3. 評価実験

提案法を検証するため，音響モデルに基づく採譜結果と音響・言語モデル統合に基づく採譜結果を比較する．

### 3.1 実験条件

RWC ポピュラー音楽データベース [5] のうち，ドラムパートを含む 4/4 拍子の 80 曲を使用し，バスドラム，スネアドラム，ハイハットの 3 種類を対象とした（$K = 3$）．MCMS のパラメータは [4] に従った（$C = 3$）．また，前処理として OpenUnmix[6] で得られたドラムパートの分離音を利用した．CNN の学習には，80 曲からランダムに選んだ 56 曲を使用し，残りをテストデータとした．VAE の学習には，J ポップとビートルズの計 534 曲のドラムパートのうち，4/4 拍子のものを用いた．

CNN と VAE の構造を図 2 に示す．CNN は各パート独立に 2 層の畳み込み層と 2 層の全結合層から構成され，畳み込み層ではバッチ正則化を行った．カーネルサイズは [4] に従い，時間方向の次元数が一定になるようにゼロパディングを施した．過学習を防ぐために，全結合層ではドロップアウト ($p = 0.2$) を適用し，$\lambda = 10^{-5}$ の重み正則化を行った．また，$a = 10$ とした．VAE のエンコーダは 2 層の全結合層，デコーダは 3 層の全結合層で構成され，最終層でドロップアウト ($p = 0.2$)，その他の層でバッチ正則化を行った．潜在空間の次元数 $D$ は 2 に設定し，各小節ごと独立に学習を行った．両ネットワークの最適化には Adam ($\mathrm{lr} = 10^{-2}$) を用いた．また，$\alpha = 0.2$，$\mathbf{Y}$ の初期化に用いる閾値は 0.5 とし，フリップは各小節ごとに 30 回繰り返した．

表 1: 評価結果

| 手法 | パート | $\mathcal{P}$(%) | $\mathcal{R}$(%) | $\mathcal{F}$(%) |
|---|---|---|---|---|
| CNN (Frame) | HH | 70.4 | 77.4 | 73.7 |
| | SD | 81.6 | 44.9 | 58.0 |
| | BD | 89.7 | 72.3 | 80.1 |
| CNN (Tatum) | HH | **67.6** | 77.0 | 72.0 |
| | SD | **74.3** | 43.8 | 55.1 |
| | BD | **83.3** | 72.6 | 77.6 |
| CNN+VAE (Tatum) | HH | 67.1 | **77.6** | 72.0 |
| | SD | 73.8 | **44.9** | **55.8** |
| | BD | 82.9 | **73.4** | **77.8** |

統合モデルの評価はテイタム単位で行い，ベースラインとなる CNN 音響モデルに対してはフレーム・テイタム単位両方の評価を行った．フレーム単位の評価指標として，発音時刻に対する適合率 ($\mathcal{P}$)，再現率 ($\mathcal{R}$) および F 値 ($\mathcal{F}$) を用いた．検出許容誤差は 50 ms とした．

### 3.2 実験結果

各パートの再現率，適合率および F 値を表 1 に示す．音響モデルのみを利用した場合 (CNN) よりも，言語モデルを統合した場合 (CNN+VAE) の方が，適合率は低下し，再現率とハイハットを除く F 値はわずかに向上したが，大きな差は認められなかった．大きな向上が得られなかった原因として，音響モデルが CNN の出力以外の楽譜に対して，不適切に低い評価値を与えてしまう問題がある．また，ランダムフリップを用いた推定ドラム譜の最適化が局所解に陥る問題の影響も考えられる．

## 4. おわりに

本稿では，音楽的な妥当性を評価する言語モデルを導入することで，音響モデルのみで推定される楽譜をより正解に近づける方法を提案した．具体的には，CNN を用いた深層音響モデルに対して，VAE に基づく言語モデルを統合する手法を考案した．今後は，本手法の各構成要素の改良に加えて，楽曲の繰り返し構造を反映させたグローバルな単位で言語モデルを定式化することで，採譜精度の向上を目指す．さらに，言語モデルを用いて音響モデルを半教師あり学習することで，データセットが不足した状況でも機能するモデルを考案する予定である．

**謝辞** 本研究の一部は，JST ACCEL No. JPMJAC1602 および科研費 No. 19H04137, No. 19K20340, No. 16H01744 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 with template adaptation. *ISMIR*, 257–263, 2015.

[2] R. Stables *et al.* Automatic drum transcription using bi-directional recurrent neural networks. *ISMIR*, 591–597, 2016.

[3] S. Ueda *et al.* Bayesian drum transcription based on non-negative matrix factor decomposition with a deep score prior. *ICASSP*, 456–460, 2019.

[4] J. Schlüter *et al.* Improved musical onset detection with convolutional neural networks. *ICASSP*, 6979–6983, 2014.

[5] M. Goto *et al.* RWC music database: Popular, classical and jazz music databases. *ISMIR*, 287–288, 2002.

[6] F.-R. Stöter *et al.* Open-Unmix - A reference implementation for music source separation. *JOSS*, 2019.



Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.