Vol.2020-MUS-128 No.8
2020/8/25

事前学習済み言語モデルによる正則化を用いた深層ニューラルネットワークに基づくドラム採譜

石塚 崚斗[1,a)]　錦見 亮[1,b)]　中村 栄太[1,c)]　吉井 和佳[1,d)]

概要：本稿では，音響音楽信号からドラムのオンセット時刻をテイタム単位で推定する手法を述べる．自動ドラム採譜では，フレーム単位で設計された深層ニューラルネットワーク (deep neural network; DNN) により，スペクトログラムを入力としてドラムのオンセット時刻を出力する手法が盛んに研究されてきた．しかし，フレーム単位の DNN では楽曲の繰り返し構造を学習することが難しいために，音楽的に不自然なドラムパターンをしばしば生成してしまうという問題点が指摘されていた．この問題を解決するため，我々はフレーム単位の入力特徴量からテイタム単位のドラム譜を推定する DNN に対して，推定結果を音楽的に妥当なパターンに誘導する正則化を施す学習手法を提案する．提案法では，大規模なドラム譜から学習された統計的言語モデルとして，GRU (gated recurrent unit) とスキップタイプ bi-gram を採用する．テイタム単位で推定されたオンセット時刻の音楽的な妥当性を事前学習済みの統計的言語モデルで評価することで，学習過程において正則化を行う．標準データセットを用いた実験により，提案法の効果を示す．

キーワード：自動ドラム採譜，言語モデル，正則化

1. はじめに

自動ドラム採譜 (automatic drum transcription; ADT) は自動音楽採譜 (automatic music transcription; AMT) のタスクであり，音楽音響信号から楽譜を推定することを目的とする．ドラムがポピュラー音楽の音楽的な構造を支える重要な楽器であることから，自動ドラム採譜は自動音楽採譜の中でも特に重要な役割を果たしている．本稿では，ドラムの中でも主要な bass drum (BD)，snare drum (SD)，hi-hats (HH) の 3 楽器を扱う．一般的には，自動ドラム採譜はフレーム単位の入力特徴量からドラムが演奏されている秒数 (オンセット時刻) の推定を行うが，記号単位の推定を行う研究はまだ少ない．

自動ドラム採譜では，DNN と非負値行列因子分解 (non-negative matrix factrization; NMF) に基づく手法が代表的であり，入力特徴量であるスペクトログラムからフレーム単位でオンセット確率値を出力する [1]．特に，畳み込みニューラルネットワーク (convolutional neural networks; CNN) は局所的な時間-周波数の領域に着目して特徴量を抽出することにより，高い精度を達成している [2–4]．さらに，再帰型ニューラルネットワーク (recurrent neural networks; RNN) を用いることで，長期的な時系列を考慮してドラム譜を推定できる [5–7]．ドラムのスペクトルは類似した波形の楽器音から構成されるため，自動ドラム採譜では依然として NMF もよく用いられる [8–12].

しかし，フレーム単位で設計された自動ドラム採譜システムでは，音楽的に不自然なパターンが出力されるのを防ぐ機構がないため，繰り返し構造に代表されるような楽曲の特徴を学習することが難しい．この問題を解決する手法の一つに，言語モデルはテイタム単位におけるオンセットの確率分布を表すことでドラム譜の音楽的な妥当性を評価する言語モデルの利用が挙げられる．実際に，ベイズ推論の枠組みで，NMF を用いた音響モデルとドラム譜の言語モデルを統合する手法も提案されている [10]. しかし，NMF の表現力が乏しいために推定精度が十分でないことや，事後分布の推論に時間を要するという問題が残されている．

言語モデルは自動音声認識 (automatic speech recognition; ASR) で重要な役割を果たしており，入力特徴量から言語的に自然な単語列を推定するために利用されている．言語モデルを利用した伝統的なアプローチでは，単語列の生成過程を表す単語単位の言語モデル (例えば n-gram) と，単語列から特徴量の生成過程を表すフレーム単位の音響モデル (例えば hidden Markov model; HMM) を統合す

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

る [13]．特徴量から単語を推定するときには，デコーダ (例えば weighted finite-state transducer; WFST) を用いて，音響面と言語面の両方を考慮する推論が行われる．この手法は，ペアデータ (音声データとアノテーション) から音響的な側面を，テキストデータから言語的な側面をそれぞれ個別に学習できるという点で優れている．

最近では，特徴量から単語列を直接推定する End-to-end 型の DNN に基づく ASR の手法が盛んに研究されている．代表的なアプローチは注意機構に基づくエンコーダ・デコーダモデルであり，エンコーダは音響的な特徴量を抽出し，デコーダは言語的な推論を行う機構とみなすことができる [14]．このモデルは実装が容易で推論時間も短いという利点がある一方で，学習に大量のペアデータが必要となる欠点がある．そこで，テキストデータを転移学習によって活用しようとする研究が提案されてきた [15, 16]．知識蒸留 [17] は転移学習の一種で，大規模なテキストデータによって学習された言語モデルを End-to-end 型モデルの学習に活用する方法の一つである．

このような背景から，我々は音響音楽信号から得られたフレーム単位のスペクトログラムからテイタム単位のオンセット時刻を推定する採譜モデルを設計し，convolutional RNN (CRNN) を用いる (図 1 青線部)．CNN がフレームレベルで特徴量を抽出する機構として働き，RNN がテイタム単位の音楽的な構造を捉える機構として働くため，CRNN はエンコーダ・デコーダモデルとして音響的な側面と言語的な側面の両者を同時に学習することが期待される．本提案法では，ビート時刻を予め推定し，max pooling を用いてフレーム・テイタム間のアライメントを行う (図 1 青線部内)．

大規模ドラム譜データを利用するため，我々は CRNN の学習過程で推定結果を音楽的に自然なパターンへと誘導する正則化法を提案する (図 1 赤線部)．具体的には，モデルが出力するテイタム単位オンセット確率値と正解ラベル間のクロスエントロピー $\mathcal{L}_{\text{ce}}$ と，オンセット確率値の負の対数尤度 $\mathcal{L}_{\text{lang}}$ との重み付け和 $\mathcal{L}_{\text{total}}$ を目的関数として最小化する．このとき，オンセット確率値は gumbel-sigmoid trick [18] により微分可能な形で二値化されるため，誤差逆伝播に基づくネットワークの最適化が可能になる．

2. 関連研究

本章では，言語モデルと知識移転に関するものを中心に，自動音楽採譜と自動ドラム採譜の関連研究を述べる．

2.1 自動ドラム採譜

NMF は，振幅スペクトログラムを基底行列とアクティベーション行列に分解する手法として自動ドラム採譜に応用されてきた [8, 11, 12]．しかし，基底行列の表現力の乏しさから，最近では局所的な特徴量を自動で抽出する

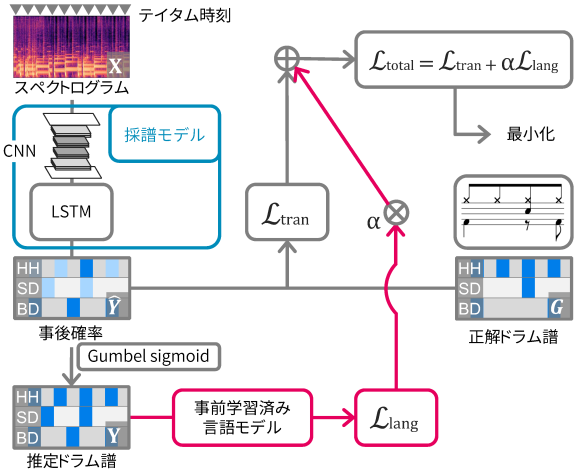

図 1　本提案法の概略図．

CNN が利用されるようになり，自動ドラム採譜だけでなく自動音楽採譜で広く活用されている [2–4, 19]．フレーム単位の長期的な時系列依存性を捉えるために，RNN もよく利用されている．Vogl らは，自動ドラム採譜に RNN を導入し [5]，ビート推定とのマルチタスク学習の効果を示した [7]．このように，自動ドラム採譜では音響音楽信号とアノテーションのペアデータにより学習された DNN ベースの手法が高い精度を発揮している．

2.2 言語モデル

自動音楽採譜のシステムの性能を改善する手法の一つとして，音楽的な妥当性を評価する言語モデルの利用が挙げられる．従来，言語モデルはフレーム単位で設計されてきた．例えば，Raczyński ら [20] は deep belief network を用いて音高と音価をモデル化した．Sigtia ら [21] は，RNN によって推定された事後確率から音楽的に妥当なコード系列を推定するために言語モデルを用いた．しかし，[22, 23] で指摘されているように，言語モデルが音楽的な構造を学習するためには，テイタム単位で設計する必要がある．

最近になって，自動音楽採譜の分野ではテイタム単位で設計された言語モデルの利用が検討されている．Wang ら [24] は，音符単位で設計された言語モデルを自動ピアノ採譜に応用した．Korzeniowski ら [25, 26] は，シンボル単位の言語モデルをコード認識に応用した．Ycart ら [27] は，LSTM の言語モデルとしての能力を幅広い観点から調査し，16 分音符単位で設計された LSTM 言語モデルが音符の遷移などを表現することができることを示した．自動ドラム採譜では，Thompson ら [28] がサポートベクターマシンを用いてドラムパターンをいくつかに分類して辞書を作り，そのテンプレートとのマッチングに基づく言語モデルを提案した．上田ら [10] は，DNN ベースの言語モデルをドラム譜の事前分布として利用することで，ベイズ推論を行う手法を提案した．しかし，テイタム単位の言語モデルを DNN ベースの自動ドラム採譜システムに組み込む試

みはまだなされていない．

2.3 転移学習

転移学習は関連するドメインから効率的に知識を利用することを目的とし，多くの分野で幅広く応用されている [29]．転移学習の扱う対象はペアデータだけでなく，ラベルなしデータに対しても適用することができ，対象とする問題に応じて様々なデータセットの組み合わせが利用される [17, 30, 31]．いくつかの研究では，teacher-student の枠組みで student モデルが teacher モデルと同等もしくは上回る性能を発揮することが報告されている [32, 33]．転移学習が比較的軽量な student モデルの学習に利用される場合は，特に知識蒸留と呼ばれる [17]．

最近，大規模ラベルなしデータから学習された他ドメインの知識を利用する研究が盛んに行われている．ASR では，言語として自然な単語列を推定するために，デコード時に言語モデルが利用されている．しかし，この手法では推論に長い時間を要する．このような背景から，ASR では知識蒸留に基づく手法が提案されている．cold fusion [34] と呼ばれる手法では，End-to-end 型モデルの学習時に事前学習済み言語モデルによる正則化項を組み込むことで転移学習を行なっている．この手法は推論時に言語モデルを必要としないため，高速に動作する利点がある [15]．知識蒸留は自動ドラム採譜でも利用されており，Wu ら [35] は DNN ベースの student モデルに対して NMF ベースの teacher モデルの知識を埋め込むことで，ラベルなしデータの活用方法を示した．

3. 提案手法

本章では，音楽音響信号から得られたメルスペクトログラムを入力として，ドラム譜を推定する提案法について，問題設定を述べる (3.1 節)．図 1 に示す通り，我々の提案法はテイタム単位のオンセット確率値を推定する DNN ベースの採譜モデル (3.2 節) を利用し，ドラム譜の音楽的な妥当性を評価する言語モデル (3.3 節) を大規模ドラム譜により事前に学習しておくことで，採譜モデルの学習時に正則化項として機能させている (3.4 節)．

3.1 問題設定

我々は，メルスペクトログラム $\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ からドラム譜 $\mathbf{Y} \in \{0,1\}^{K \times M}$ を推定する問題に取り組む．ここで，K は扱うドラムの楽器数 (本稿では BD，SD，HH であるため $K = 3$)，F は周波数ビン数，T は時間フレーム数を表す．本稿では，全てのオンセット時刻がビート時刻を四等分したテイタム時刻に沿っているものと仮定し，ビート時刻は予め推定された結果を利用する．

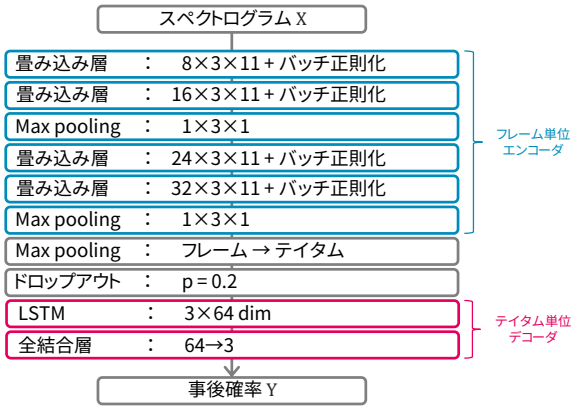

図 2 CRNN の詳細図．モデルはフレーム単位のエンコーダとテイタム単位のデコーダから構成され，max pooling を利用してフレーム・テイタム間のアライメントをとる．エンコーダは畳み込み層 4 層と max pooling 2 層，デコーダは GRU 3 層と全結合層 1 層から構成される．

3.2 採譜モデル

採譜モデルは，テイタム単位のオンセット確率値 $\tilde{\mathbf{Y}} \in [0,1]^{K \times M}$ を推定する．ここで，$\tilde{Y}_{km}$ はドラム k のテイタム m におけるオンセット確率値を表す．学習時には，正解ドラム譜 $\mathbf{G} \in \{0,1\}^{K \times M}$ が与えられた下で，$\tilde{\mathbf{Y}}$ と $\mathbf{G}$ のクロスエントロピー $\mathcal{L}_{\text{tran}}(\tilde{\mathbf{Y}}|\mathbf{G})$ を最小化することでネットワークを最適化する．

$$\mathcal{L}_{\text{tran}}(\tilde{\mathbf{Y}}|\mathbf{G}) = -\sum_{k=1}^{K}\sum_{m=1}^{M}\big(G_{km}\log\tilde{Y}_{km} + (1-G_{km})\log(1-\tilde{Y}_{km})\big). \quad (1)$$

採譜時には，$\tilde{\mathbf{Y}}$ を閾値 $\delta \in [0,1]$ によって二値化することで，ドラム譜 $\mathbf{Y} \in \{0,1\}$ を推定する．採譜モデルの実装には，CRNN を用いる (図 2)．最初に，CRNN は $\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ を入力として受け取り，フレーム単位の潜在特徴量 $\mathbf{F} \in \mathbb{R}^{D \times T}$ に変換する．ここで，D は潜在空間の次元数とする．次に，テイタム時刻 $\mathbf{B} = \{b_m\}_{m=1}^{M}$ に基づく max pooling を利用して，フレーム単位の潜在特徴量 $\mathbf{F} \in \mathbb{R}^{D \times T}$ をテイタム単位の潜在特徴量 $\tilde{\mathbf{F}} \in \mathbb{R}^{D \times M}$ に変換する．具体的には，以下の操作を行う．

$$\tilde{\mathbf{F}}_{d,m} = \max\left\{F_{d,b_{m-1}:b_m}\right\} \quad (2)$$

ここで，“i:j” はインデックス i からインデックス j までの要素を表し，$b_0 = 0$ とする．最後に，GRU と全結合層を通してオンセット確率値 $\tilde{\mathbf{Y}}$ を推定する．

3.3 言語モデル

言語モデルは，ドラム譜 $\mathbf{Y}$ の生成確率を評価する．学習時には，負の対数尤度 $\mathcal{L}_{\text{lang}}(\mathbf{Y}) = -\log p(\mathbf{Y})$ を最小化する．

$$\mathcal{L}_{\text{lang}}(\mathbf{Y}) = -\sum_{m=1}^{M}\log p(Y_{:,m}|Y_{:,1:m-1}). \quad (3)$$

ここで，“:” は全ての要素を表す．

言語モデルの実装には，単純な構造でありながらも強力なモデルであるスキップタイプ bi-gram と，LSTM と比べて同等の性能を有しながらも学習が容易である双方向 GRU を用いた．スキップタイプ bi-gram では，多くのポピュラー音楽は 4/4 拍子で構成され，類似したドラムパターンが現れやすいという性質から，ドラム譜 $\mathbf{Y}$ の小節ごとの繰り返し構造を以下のようにモデル化する．

$$p(Y_{:,m}|Y_{:,1:m-1}) = \prod_{k=1}^{K} p(Y_{k,m}|Y_{k,m-16}), \\ = \prod_{k=1}^{K} \pi_{Y_{k,m-16},Y_{k,m}}. \quad (4)$$

ここで，$\pi_{A,B}$ $(A, B \in \{0, 1\})$ は A から B への遷移確率を表している．学習時には，最尤推定を用いて遷移確率 $\pi_{A,B}$ $(A, B \in \{0, 1\})$ を求める．一方で，GRU を用いた言語モデルでは，$p(Y_{:,m}|Y_{:,1:m-1})$ を直接学習し，三楽器の相互依存性をモデル化する．

3.4 正則化

我々は，以下で表される誤差関数 $\mathcal{L}_{\text{total}}$ を最小化することで誤差逆伝播に基づいて採譜モデルの最適化を行う．

$$\mathcal{L}_{\text{total}}(\tilde{\mathbf{Y}}, \mathbf{Y}|\mathbf{G}) = \mathcal{L}_{\text{tran}}(\tilde{\mathbf{Y}}|\mathbf{G}) + \alpha\mathcal{L}_{\text{lang}}(\mathbf{Y}). \quad (5)$$

ここで，$\alpha > 0$ は事前学習済み言語モデルによる正則化項の重みである．学習時には，$\tilde{\mathbf{Y}}$ を閾値によって二値化する代わりに，誤差逆伝播を行うために微分可能な形で二値化する．そこで，我々は gumbel-sigmoid trick [18] と呼ばれる手法を用いる．

$$U_{km}^{(i)} \sim \text{Uniform}(0, 1), \quad (6)$$

$$V_{km}^{(i)} = -\log\left\{-\log\left(U_{km}^{(i)}\right)\right\}, \quad (7)$$

$$Y_{km} = \sigma\left\{\frac{\tilde{Y}_{km} + V_{km}^{(1)} - V_{km}^{(2)}}{\tau}\right\}. \quad (8)$$

ここで，$i = 1, 2$, $\tau > 0$ は温度パラメータ，$\sigma(\cdot)$ はシグモイド関数を表す．事前学習済み言語モデル (スキップタイプ Bi-gram もしくは GRU) のパラメータは，採譜モデルの学習時は固定しておく．

4. 評価実験

本章では，提案法に基づく学習法の有効性と精度評価の結果を報告する．

4.1 実験設定

採譜モデルの評価には RWC ポピュラー音楽データベース (RWC) と ENST-drums (ENST) を利用した．RWC ではドラムが含まれる計 89 曲，ENST では wet-mix の minus-one track 計 64 曲を利用した．RWC はランダム，ENST はドラマーごとに 3 分割して 3-交差分割検証を行なった．それぞれの訓練データのうち，ランダムに 15% を選んで検証データとして利用した．それぞれの楽曲を 44.1kHz でリサンプリングし，シフト幅 441 フレーム (10ms)，窓幅 2048 フレームの短時間フーリエ変換を用いて振幅スペクトログラムを作成した．さらに，バンド数 80，最低周波数 20Hz，最高周波数 20000Hz のメルフィルタバンクを利用して，入力特徴量のメル周波数スペクトログラムを作成した．ビート推定には madmom [36] を利用した．ただし，RWC のビート推定は Open-Unmix [37] を用いて前処理を行なった音源に対して行なった．

表 1 ビート推定の精度 (%)，検出不可能なオンセット (%)，事前学習済み言語モデルによるテストセットパープレキシティ．

dataset	ビート	オンセット			パープレキシティ	
	$\mathcal{F}$	*double*	*far*	*comb*	Bi-gram	GRU
RWC	84.6	0.33	0.61	0.92	1.48	1.00
ENST	n/a	0.51	0.22	0.71	1.78	1.00

CRNN の構造は [38] に基づいて設計した．畳み込み層のカーネルサイズは 3×11 であり，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．カーネルサイズは [2, 38] で提案されているサイズより時間方向に長いが，これは [7] で示されている知見に基づいて設定している．両ネットワークの最適化には Adam (lr=10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$) を利用し，検証データの目的関数が 10 回連続で下がらなかった場合に学習を停止させた．その後，学習率を各エポックごとに 0.9 倍して同様の学習を行なった．過学習を防ぐために重み正則化 (λ=10^{-4})，全結合層の直前にドロップアウト ($p = 0.2$) を適用した．ハイパーパラメータは，a=3, $\tau = 0.3$, $\delta = 0.2$ に設定した．言語モデルの学習には Jpop とビートルズの計 534 曲のドラム譜を用い，スイップタイプ bi-gram の遷移確率と双方向 GRU を学習した．なお，GRU の層数は 3 に設定した．言語モデルの重み α は $(0, 2]$ の範囲でグリッドサーチを行なった．

評価尺度には，以下で定義される F 値を利用した．

$$\mathcal{P} = \frac{N_c}{N_e}, \quad \mathcal{R} = \frac{N_c}{N_g}, \quad \mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}}. \quad (9)$$

ここで，N_e は推定されたオンセット・ビート数，N_g は正解のオンセット・ビート数，N_c は正解したオンセット・ビート数を表している．推定されたオンセット・ビート時刻には，50ms の許容誤差を設定し，mir_eval [39] を利用して $\mathcal{P}$, $\mathcal{R}$, $\mathcal{F}$ を計算した．

4.2 データセットの統計的性質

RWC におけるビート推定の精度を表 1 の左側に示す．ただし，ENST にはビートアノテーションが付属していな

いため，精度評価を行なっていない．この結果から，RWC では比較的高い精度でビートを推定できていることが分かる．

我々の提案法では，オンセット時刻がテイタム時刻に一致していると仮定しているため，検出することができないオンセットが存在する．*double* は 1 つのテイタムグリッドの中に 2 つ以上存在しているオンセットを表し，*far* はオンセット時刻がテイタム時刻から許容誤差 (本稿では 50ms) 以内に位置しないオンセットを表す．結局，本提案法を利用した場合に検出することができないオンセットは，以下のように表される．

$$N_{comb} = N_{double} + N_{far} - N_{\{double \cap far\}}, \qquad (10)$$

ここで，N_{double} は *double* に属するオンセット数，N_{far} は *far* に属するオンセット数，$N_{double \cap far}$ は *double* と *far* に属するオンセット数を表す．表 1 の中央は全体のオンセット数における N_{double}，N_{double}，$N_{double \cap far}$ の割合を示している．この結果から，我々の提案法では 2 つのデータセットに含まれるほとんどのオンセットを検出できることが分かる．

それぞれのデータセットに対して，事前学習済み言語モデルに基づいてテストセットパープレキシティを計算した結果を表 1 の右側に示す．この結果から，スキップタイプ bi-gram に比べて GRU の方が，評価データに対する言語モデルとして適切に機能していることが分かる．

4.3 実験結果

提案法を利用して学習を行なった CRNN の採譜精度を表 2 に示す．事前学習済み言語モデルによる正則化が有効であることが分かった．また，言語モデルには，スキップタイプ bi-gram を用いるよりも GRU を用いた方が高い精度を示している．従来法 [7] と比べて精度が劣っているが，これは我々の提案法における推定オンセット時刻があらかじめ推定されたビート時刻に沿っているという仮定に起因しているものと考えられる．

本提案法の正則化に基づく学習法によって，推定結果が改善した例を図 3，4 に示す．両図とも，最上段は正解ドラムパターン，中段は正則化を施さない採譜モデルの推定結果，最下段は正則化を施した採譜モデルの推定結果を表す．いずれの例においても，正則化を行わない採譜モデルでは音楽的に不自然なドラムパターン (赤枠線部) を推定していたが，正則化を行うことで自然なドラムパターンを推定していることが分かる．同時に，正則化を行うことで余分なオンセット (青枠線部) を推定する場合や，本来推定できていたオンセット (青枠点線部) を推定できなくなってしまう場合がある．

表 2 提案法の精度．α の左側は RWC, 右側は ENST に対する言語モデルの重みを表している．

	RWC			ENST		
	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
CRNN [7]	n/a	n/a	n/a	78.4	n/a	n/a
CRNN	62.5	52.0	79.9	60.9	52.0	74.2
+ Bi-gram ($\alpha = 0.50, 0.18$)	64.7	56.8	77.5	62.0	55.6	72.3
+ GRU ($\alpha = 0.14, 0.07$)	**67.6**	66.8	70.7	**62.1**	56.8	70.5

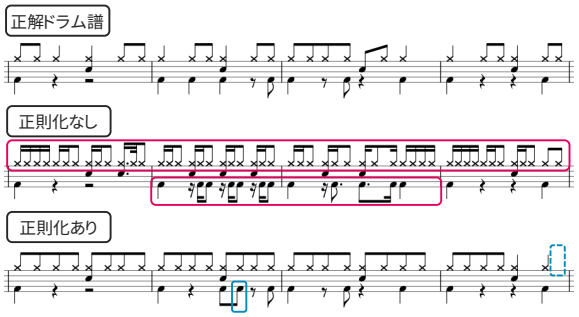

図 3 採譜モデルに正則化を施すことで，推定結果が改善した具体例 (RWC-MDB-P-2001 No.14 の一部)．

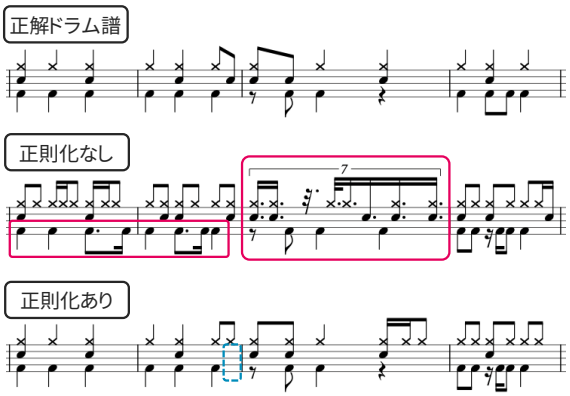

図 4 採譜モデルに正則化を施すことで，推定結果が改善した具体例 (RWC-MDB-P-2001 No.26 の一部)．

5. おわりに

本稿では，事前学習済み言語モデルによる正則化を用いた CRNN の学習法を提案した．CRNN は，フレーム単位の特徴量抽出器であるエンコーダとテイタム単位で音符の時系列依存性を表現するデコーダで構成され，音響的なモデルと言語的なモデルを同時に学習する．評価実験により，事前学習済み言語モデルによる正則化は，採譜モデルの精度と音楽的な自然さを向上させることが明らかになった．我々の提案法では従来法の精度に及ばなかったが，ビート推定の精度が影響を与えていることが考えられる．

今後は，自動ドラム採譜の目的が記号単位の楽譜を出力することであるため，従来のようなフレーム単位の評価ではなく記号単位の評価尺度を導入する予定である．他にも，ドラムのオンセット時刻とビート時刻に相互依存性が

あるとすれば，マルチタスク学習の枠組みでビート・ダウンビート推定を我々の提案法に組み入れることが考えられる．さらに，フレーム単位からドラム譜を推定する発展的な手法として，採譜モデルと言語モデルに self-attention を利用することで，よりグローバルにドラムパターンの構造を捉えることが考えられる．

謝辞 本研究の一部は，JST ACCEL No. JPMJAC1602，JSPS 科研費 No. 16H01744，19K20340 および 19H04137 の支援を受けた．

参考文献

[1] Wu, C.-W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Müller, M. and Lerch, A.: A review of automatic drum transcription, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 9, pp. 1457–1483 (2018).

[2] Jacques, C. and Roebel, A.: Automatic drum transcription with convolutional neural networks., *DAFx*, pp. 80–86 (2018).

[3] Gajhede, N., Beck, O. and Purwins, H.: Convolutional neural networks with batch normalization for classifying hi-hat, snare, and bass percussion sound samples., *Proceedings of the Audio Mostly 2016*, pp. 111–115 (2016).

[4] Southall, C., Stables, R. and Hockman, J.: Automatic Drum Transcription for Polyphonic Recordings Using Soft Attention Mechanisms and Convolutional Neural Networks., *ISMIR*, pp. 606–612 (2017).

[5] Vogl, R., Dorfer, M. and Knees, P.: Recurrent Neural Networks for Drum Transcription., *ISMIR*, pp. 730–736 (2016).

[6] Stables, R., Hockman, J. and Southall, C.: Automatic Drum Transcription using Bi-directional Recurrent Neural Networks., *ISMIR*, pp. 591–597 (2016).

[7] Vogl, R., Dorfer, M., Widmer, G. and Knees, P.: Drum Transcription via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks., *ISMIR*, pp. 150–157 (2017).

[8] Wu, C.-W. and Lerch, A.: Drum Transcription Using Partially Fixed Non-Negative Matrix Factorization with Template Adaptation., *ISMIR*, pp. 257–263 (2015).

[9] Roebel, A., Pons, J., Liuni, M. and Lagrangey, M.: On automatic drum transcription using non-negative matrix deconvolution and itakura saito divergence., *ICASSP*, pp. 414–418 (2015).

[10] Ueda, S., Shibata, K., Wada, Y., Nishikimi, R., Nakamura, E. and Yoshii, K.: Bayesian Drum Transcription Based on Nonnegative Matrix Factor Decomposition with a Deep Score Prior., *ICASSP*, pp. 456–460 (2019).

[11] Dittmar, C. and Gärtner, D.: Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition., *DAFx*, pp. 187–194 (2014).

[12] Paulus, J. and Klapuri, A.: Drum sound detection in polyphonic music with hidden markov models., *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2009, pp. 1–9 (2009).

[13] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups., *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97 (2012).

[14] Chorowski, J., Bahdanau, D., Cho, K. and Bengio, Y.: End-to-end continuous speech recognition using attention-based recurrent NN: first results., *NIPS Workshop on Deep Learning* (2014).

[15] Kubin, G. and Kacic, Z.: Learn Spelling from Teachers: Transferring Knowledge from Language Models to Sequence-to-Sequence Speech Recognition., *Interspeech* (2019).

[16] Chen, Y.-C., Gan, Z., Cheng, Y., Liu, J. and Liu, J.: Distilling the Knowledge of BERT for Text Generation., *arXiv* (2019).

[17] Hinton, G., Vinyals, O. and Dean, J.: Distilling the knowledge in a neural network., *arXiv* (2015).

[18] Tsai, Y.-H., Liu, M.-Y., Sun, D., Yang, M.-H. and Kautz, J.: Learning binary residual representations for domain-specific video streaming., *AAAI*, pp. 7363–7370 (2018).

[19] Schlüter, J. and Böck, S.: Improved musical onset detection with convolutional neural networks., *ICASSP*, pp. 6979–6983 (2014).

[20] Raczyński, S. A., Vincent, E. and Sagayama, S.: Dynamic Bayesian networks for symbolic polyphonic pitch modeling., *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 9, pp. 1830–1840 (2013).

[21] Sigtia, S., Boulanger-Lewandowski, N. and Dixon, S.: Audio Chord Recognition with a Hybrid Recurrent Neural Network., *ISMIR*, pp. 127–133 (2015).

[22] Korzeniowski, F. and Widmer, G.: On the futility of learning complex frame-level language models for chord recognition., *ISMIR*, pp. 10–17 (2017).

[23] Ycart, A., McLeod, A., Benetos, E., Yoshii, K. et al.: Blending acoustic and language model predictions for automatic music transcription., *ISMIR*, pp. 454–461 (2019).

[24] Wang, Q., Zhou, R. and Yan, Y.: A two-stage approach to note-level transcription of a specific piano., *Applied Sciences*, Vol. 7, No. 9, p. 901 (2017).

[25] Korzeniowski, F. and Widnaer, G.: Automatic Chord Recognition with Higher-Order Harmonic Language Modelling., *EUSIPCO*, pp. 1900–1904 (2018).

[26] Korzeniowski, F. and Widmer, G.: Improved chord recognition by combining duration and harmonic language models., *ISMIR*, pp. 10–17 (2018).

[27] Ycart, A., Benetos, E. et al.: A study on LSTM networks for polyphonic music sequence modelling., *ISMIR*, pp. 421–427 (2017).

[28] Thompson, L., Mauch, M., Dixon, S. et al.: Drum transcription via classification of bar-level rhythmic patterns., *ISMIR*, pp. 187–192 (2014).

[29] Weiss, K., Khoshgoftaar, T. M. and Wang, D.: A survey of transfer learning., *Journal of Big data*, Vol. 3, No. 1, p. 9 (2016).

[30] Zagoruyko, S. and Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer., *ICLR* (2017).

[31] Yim, J., Joo, D., Bae, J. and Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning., *CVPR*, pp. 4133–4141 (2017).

[32] Mobahi, H., Farajtabar, M. and Bartlett, P. L.: Self-distillation amplifies regularization in hilbert space., *arXiv* (2020).

[33] Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L. and Anandkumar, A.: Born again neural networks., *ICML*,

pp. 1602–1611 (2018).
[34] Sriram, A., Jun, H., Satheesh, S. and Coates, A.: Cold fusion: Training seq2seq models together with language models., *Interspeech*, pp. 387–391 (2018).
[35] Wu, C.-W. and Lerch, A.: Automatic Drum Transcription Using the Student-Teacher Learning Paradigm with Unlabeled Music Data., *ISMIR*, pp. 613–620 (2017).
[36] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F. and Widmer, G.: Madmom: A new python audio and music signal processing library., *ACM international conference on Multimedia*, pp. 1174–1178 (2016).
[37] Stöter, F., Uhlich, S., Liutkus, A. and Mitsufuji, Y.: Open-Unmix - A Reference Implementation for Music Source Separation, *J. Open Source Softw.*, Vol. 4, No. 41, p. 1667 (2019).
[38] Vogl, R., Widmer, G. and Knees, P.: Towards multi-instrument drum transcription., *DAFx*, pp. 57–64 (2018).
[39] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., Ellis, D. P. and Raffel, C. C.: mir_eval: A transparent implementation of common MIR metrics., *ISMIR*, pp. 367–372 (2014).

5 ページ，左カラム，29-34 行目	また，言語モデルには，スキップタイプ bi-gram を用いるよりも GRU を用いた方が高い精度を示している．従来法[7]と比べて精度が劣ってるが，これは我々の提案法における推定オンセット時刻があらかじめ推定されたビート時刻に沿っているという仮定に起因しているものと考えられる．	また，言語モデルには，GRU を用いるよりもスキップ型 bi-gram を用いた方が高い精度を示している．これは，4.2 節で述べた実験結果と矛盾しており，今後精査する必要がある．言語モデルの最適な重みはデータセットとモデルの両方に依存することが分かる．ENST に対する - の評価値では，従来法 [7]と比べて精度が劣っているが，4.2 節で述べた通り我々の提案法では原理的に検出できないオンセットの割合は非常に小さいため，さらなる性能改善の余地があると考えられる．
5 ページ，図 3		
5 ページ，図 3，キャプション	採譜モデルに正則化を施すことで，推定結果が改善した具体例 (RWC-MDB-P-2001 No.14 の一部)．	採譜モデルに正則化を施すことで，推定結果が改善した具体例 (RWC-MDB-P-2001 No.25 の一部)．言語モデルには bi-gram (. =1.2)を利用した．
5 ページ 図 4		
5 ページ，図 4，キャプション	採譜モデルに正則化を施すことで，推定結果が改善した具体例 (RWC-MDB-P-2001 No.26 の一部)．	採譜モデルに正則化を施すことで，推定結果が改善した具体例 (RWC-MDB-P-2001 No.25 の一部)．言語モデルには GRU (. =0.13)を利用した．

5 ページ，右カラム，8-9 行目	我々の提案法では従来法の精度に及ばなかったが，ビート推定の精度が影響を与えていることが考えられる．	提案法は従来法の精度に及ばなかったが，原理的には検出できないオンセットの割合は非常に小さいため，さらなる性能改善の余地があると考えられる．
参考文献		[40] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases., ISMIR, pp. 287–288 (2002). [41] Gillet, O. and Richard, G.: ENST-Drums: an extensive audio-visual database for drum signals processing., ISMIR, pp. 156–159 (2006).