Vol.2023-MUS-136 No.15
2023/2/28

定テンポ制約付きCTCに基づく自動ドラム採譜

鎌倉 大地[1,a)]　中村 栄太[1,b)]　吉井 和佳[1,c)]

概要：本稿では，ポピュラー音楽の音楽音響信号から直接ドラム譜を推定する深層ドラム採譜について述べる．ドラム採譜では，従来，深層ニューラルネットワーク (DNN) を用いて音楽音響信号中からドラムの発音時刻を検出したのち，別途推定したビート・ダウンビート情報を用いて楽譜上の位置を推定するアプローチが一般的であった．しかし，DNN の学習に必要なドラムの発音時刻が付与された音響信号は限られていた．本研究では，End-to-End 音声認識に着想を得て，時間同期は取れていないものの，インターネットから比較的容易に入手可能な楽譜（ドラムパート譜）を用いた End-to-End ドラム採譜に取り組む．具体的には，コネクショニスト時間分類 (connectionist temporal classification; CTC) に基づく損失関数を用いて，フレーム単位の音響特徴量系列をテイタム単位のドラムラベル系列に変換する DNN を学習する．ただし，通常の CTC では，自由な時間伸縮を許容した入出力系列のアラインメントを行うため，各テイタムに対応するフレーム数が不自然に大きく変動しうる問題がある．特に，ドラムが存在しないテイタムは，本来対応するフレームの音響特徴量はドラム由来のものではないので，正しくアラインメントが行えない．この問題を解決するため，テンポがほぼ一定となるアラインメントのみを考慮した CTC 損失関数を提案する．提案法の動作を検証するため，1 小節分の音楽音響信号を用いた学習・評価を行い，提案法がドラムが存在しないテイタムを含めて，ドラムラベル系列および入力とのアラインメントを正しく推定できることを確認した．今後，実際的な長さの音響信号を扱うため，計算コストの削減に取り組む．

1. はじめに

自動ドラム採譜 (automatic drum transcription; ADT) は，音楽音響信号を入力とし，ドラム譜を推定するタスクである．ポピュラー音楽において，ドラムは楽曲中の音楽的な構造を支える重要な構成要素であり，自動ドラム採譜は音楽情報処理の中でも重要な役割を果たす．高精度なドラム採譜の実現により，音楽活動の支援や音楽構造の理解への貢献が可能となる．

自動採譜手法には，MIDI 推定を行うものと楽譜推定を行うものがある．MIDI 推定は，各音符の音高と，秒単位の発音時刻と消音時刻を推定するもので，従来の多くの採譜研究ではこの問題を扱っている [1–3]．楽譜推定は，各音符の音高と，楽譜上での発音時刻と消音時刻を推定するもので，小節線や拍の位置を含むリズムの認識が必要になる．楽譜推定の方法として，MIDI 推定の結果に対してリズム認識を行う多段処理方式の方法があり [4]，ピアノ採譜では高精度の結果が報告されている [5]．また，ビート推定 [6, 7] の結果を用いて，リズムの量子化を行う方法もあり，ドラム採譜にも応用されている [8]．一方，音響信号から楽譜を直接推定する End-to-End 自動採譜の研究も行われている [9–12]．この方法は，多段処理による誤りの伝搬を防ぐとともに，音符単位で正確にアラインメントが取れた学習データを用いずとも，音響データと楽譜データのみによって学習が行えるという長所もある．

End-to-End 自動採譜を実現するには，End-to-End 音声認識と同様に，コネクショニスト時間分類 (connectionist temporal classification; CTC) [13] が有望である．すでに，音響信号の特徴量系列から，音高と音価からなる音符系列を直接推定する試みがなされている [10, 14]．この方式では，得られる音符系列の拍節構造の一貫性が担保されないという問題がある [15]．

End-to-End ドラム採譜において，この問題を回避するには，テイタムと呼ぶ拍節位置の最小単位（本研究では 16 分音符）ごとに，各ドラムの打音の有無を表すラベル（「ドラムラベル」と呼ぶ）を出力していけばよい．ただし，この方式では，ドラムが存在しないテイタムでは，無音を表す特別な記号を出力する必要が生じる．通常の CTC を用いた学習では，正解出力系列中の各テイタムは，その順番を保持しながら識別が容易な特徴量を持つ少数のフレームに対応付けられ，それ以外のフレームには特殊なブランク

[1] 京都大学大学院情報学研究科
[a)] kamakura@sap.ist.i.kyoto-u.ac.jp
[b)] eita.nakamura@i.kyoto-u.ac.jp
[c)] yoshii@i.kyoto-u.ac.jp

記号が割り当てられる．したがって，ドラムが存在するテイタムは，そのドラムに対応する特徴量をもつフレームと対応付けできる．一方，ドラムが存在しないテイタムをドラムが存在「しない」ことを示す特徴量をもつフレームと対応付けるのは原理的に困難である．なぜなら，そのようなフレームは本来，ブランク記号に割り付けるべきものであるからである．すなわち，CTC では，特定のパターンを持つ特徴量の出現があって初めてイベントの存在を認識できるのであって，特徴量の変化を伴わない「非生起」イベントをイベントとして認識することはできない．

この問題に対し，本研究では，CTC に基づく学習における入出力系列のアラインメントの際に，各テイタムが対応するフレーム数が，当該テイタムでのドラムの有無に関わらずほぼ一定となるような制約を導入する（図 1）．これは，ポピュラー音楽の場合，楽曲中でのテンポはほぼ一定に保たれることが多いため，連続するテイタムの継続時間（対応するフレームの個数）が大きく変化しないという前提に基づく．ドラムが存在するテイタムが，本来ドラムが存在する区間を超えて不自然に広範囲のフレームに対応付けられることを回避することにより，ドラムが存在しないテイタムも，ドラムが存在するテイタムとほぼ同じ長さの区間に対応付けることができる．

技術的には，CTC に基づく損失関数の計算時に，正解出力系列における隣り合うテイタム間の継続時間の遷移確率 [6,15] を考慮することで，継続時間が一定となるアラインメントパスほど損失が小さくなるよう重み付けを行う．可能なすべてのパスに関する損失の重み付き和の計算では，通常の CTC が，隠れマルコフモデル (HMM) の前向きアルゴリズムに類似した動的計画法を用いるのに対し，提案する定テンポ制約 CTC は，隠れセミマルコフモデル (HSMM) に対応するものが得られる．このとき，考慮すべきパスの数が大幅に増加することで，誤差逆伝播法に必要となる計算グラフの構築には莫大なメモリを要し，計算時間が現実的ではなくなる問題がある．本稿では，定テンポ制約 CTC の基本的な効果を実験的に検証した結果を報告し，計算量の削減については今後の課題とする．

2. 関連研究

本章では，自動ドラム採譜と End-to-End 自動採譜を中心に関連研究を述べる．

2.1 自動ドラム採譜

DNN に基づくドラム採譜の試みは数多く行われてきた [2,3,8,16–20]．標準的に，楽曲の音響信号のスペクトログラムを入力とし，ドラムの打音時刻をアノテーションしたものを訓練データとする．これは，2 次元データに変換することで，特徴量を抽出しやすいことが理由である．特徴抽出のために，畳み込み演算をするネットワークを使用

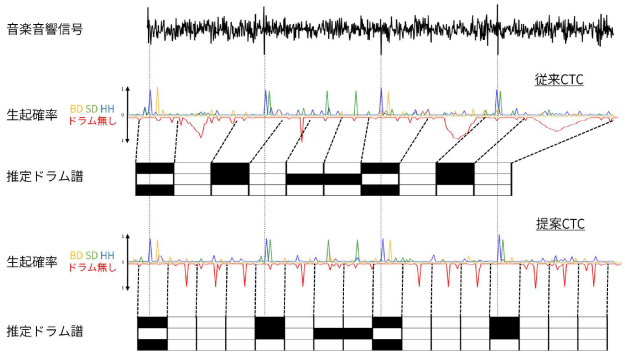

図 1 各テイタムに対応するフレーム数が一定となるような制約を導入した提案法．

することが多く，自動採譜タスクでは畳み込みニューラルネットワーク (convolutional neural network; CNN) や時間畳み込みネットワーク (temporal convolutional network; TCN) [21,22] を用いることができる．また，音楽的に意味のあるドラム譜の事前情報による正則化 [8] やネットワークへの工夫 [23] が試されており，採譜精度の向上が報告されている．これらの自動ドラム採譜は，訓練データとして楽曲のスペクトログラムとドラムの打音時刻系列が必要であり，データ量が少ないという問題があった．これに対して，合成データセットの使用 [24] やデータ拡張 [17]，教師なし学習 [25] などの手法が提案されている．しかし，これらの手法には実際の生演奏に対する頑健性やドラム音色の差異への対応能力に課題が残っている．

2.2 自動採譜における End-to-End 学習

自動採譜では，時間量子化された楽譜や楽譜上の記号を直接推定する研究も存在する [26–28]．これらは注意機構や CTC に基づく手法を採用している．例えば，歌声採譜では，テンポが一定である楽曲に対して，注意重みのピークは昇順かつ一定間隔に並ぶよう制限を加えた注意機構 [28] が提案されている．また，CTC についても，ラティス上で足し合わせるパスを制限するモデル [29] が提案されている．ここで，足し合わせるパスを一定の傾きのものだけに制限すれば，先述の制限付き注意機構と同等の効果が得られる．ドラム採譜では，定テンポ制約付きの CTC に基づく End-to-End 学習はいまだ報告されていない．

3. 提案手法

本章では，継続時間を導入した CTC を用いてドラム採譜を行う方法を説明する．

3.1 問題設定

本研究では，音楽音響信号から，ドラム譜を推定するタスクに取り組む．音響信号の左右チャネルのパワースペクトログラム $\mathbf{X} \in \mathbb{R}^{2\times F\times T}$ に対し，各テイタムにおけるドラムの有無を表すクラス $\mathbf{Y} \in \{1,2,3,\ldots,2^K\}^L$ を推定す

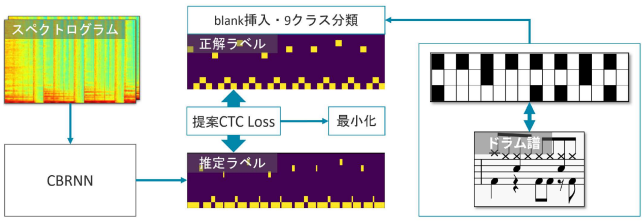

図 2 ドラム採譜のための提案 CTC に基づく CBRNN モデル.

る．ここで，F は周波数ビン数，T はフレーム数，L は対応するテイタム数，K はドラムの種類数である．本稿では，ドラムの中でも特に重要な役割を果たす 3 楽器（バスドラム (BD)・スネアドラム (SD)・ハイハット (HH)）を扱うため，$K = 3$ となる．フレームは入力するパワースペクトログラムの時間分解能を表し，10 [ms] とする．また，テイタムは楽譜上の位置を表す単位であり，本稿では 16 分音符に相当する長さを表す．従って，フレーム長は楽曲を通して一定であるが，テイタム長はテンポに依存する．

採譜システムの出力 $\mathbf{Y}$ は図 2 の右側ように，元のドラム譜から一意に定められるテイタム単位でのドラム譜を 2^K クラス分類したものである．つまり，n 番目のテイタムでのクラスは，$Y_n = 1 + 2^0\langle \mathtt{BD} \rangle_n + 2^1\langle \mathtt{SD} \rangle_n + 2^2\langle \mathtt{HH} \rangle_n$ のように表す．ただし，$\langle \mathtt{inst} \rangle_n$ は n 番目のテイタムで inst が存在するときに 1，そうでないときに 0 をとる．

3.2 CTC に基づくドラム採譜

CTC を用いた End-to-End ドラム採譜の基本的な流れについて説明する（図 2）.

3.2.1 学習

$\hat{L}$ を正解ラベル列のテイタム数としたとき, $\hat{L} < T$ である．blank ラベルを含む冗長なラベル列 $\boldsymbol{\pi} \in \{0, 1, \ldots, 2^K\}^T$ を導入することで, 入力 $\mathbf{X}$ と正解ラベル列 $\boldsymbol{l} \in \{1, 2, \ldots, 2^K\}^{\hat{L}}$ に対して，X_t に対応するラベル π_t を考えることができる．ただし，ラベル列中の 0 は blank を表し，blank は X_t が正解ラベルに含まれるどのクラスにも対応していないことを表すクラスである．

DNN は，$\mathbf{X}$ を入力として，blank ラベルを含む各フレームにおけるドラムの存在の有無を表すクラス分類確率 $\boldsymbol{\phi} \in [0, 1]^{(2^K+1)\times T}$ を出力する．訓練データである正解ラベル列を $\boldsymbol{l} \in \{1, 2, 3, \ldots, 2^K\}^{\hat{L}}$ とするとき，$\boldsymbol{\phi}$ に対して CTC に基づく誤差関数 $\mathcal{L}$ が計算できる．

$$\mathcal{L} = -\log p(\boldsymbol{l}|\mathbf{X}, \theta) \tag{1}$$

ただし，$p(\boldsymbol{l}|\mathbf{X}, \theta)$ は $\boldsymbol{\phi}$ が正解に近いほど大きくなる（詳細は 3.3）．また，θ は DNN パラメータである．DNN の学習では，$\mathcal{L}$ を最小化することが基本となる．

3.2.2 推論

$\mathbf{X}$ を入力したとき, DNN は $\boldsymbol{\phi}$ を出力する．$\boldsymbol{\phi}$ の各フレームについて, 最大となるインデックスを出力クラスとすれば, フレーム単位での出力クラス系列 $\mathbf{H} \in \{0, 1, 2, \ldots, 2^K\}^T$ は次式で決定される．

$$\mathrm{H}_t = \underset{k}{\operatorname{argmax}}\, \phi_{k,t} \tag{2}$$

さらに，最終的なテイタム単位での出力クラス $\mathbf{Y} \in \{1, 2, 3, \ldots, 2^K\}^L$ を，以下のように定める．

$$\mathbf{Y} = \mathcal{B}(\mathbf{H}) \tag{3}$$

提案法では，$\mathbf{H}$ と $\mathbf{Y}$ のアラインメントが一定の傾きで，テンポの一貫性が考慮されていることが期待される．図 3 にアラインメントの例を示す．この例では，アラインメントが直線的でないから，テンポに一貫性があるとは言えない．

3.3 通常の CTC に基づく損失関数

本稿では以降，ラベル列 $\boldsymbol{\pi}$ をパスと呼ぶこととする．パス $\boldsymbol{\pi}$ をラベル列 $\mathbf{Y} \in \{1, 2, 3, \ldots, 2^K\}^L$ に変換する関数を $\mathcal{B} : \mathbb{Z}^T \to \mathbb{Z}^L$ とし，以下の操作で定義する．

(1) 同じラベルの繰り返しを削除する
(2) 0 を全て削除する

例えば，$\mathcal{B}(10120) = \mathcal{B}(01100122) = 112$ となる．ここで，あるラベル列 $\boldsymbol{l}$ になる全てのパスの集合を $\mathcal{B}^{-1}(\boldsymbol{l})$ で表すこととすると，DNN の出力が $\mathcal{B}^{-1}(\boldsymbol{l})$ のいずれかに対応すれば，推定が正しいと見なせる．

パス $\boldsymbol{\pi}$ の確率は以下で計算できる．

$$p(\boldsymbol{\pi}|\mathbf{X}, \theta) = \prod_{t=1}^{T} \phi_{\pi_t, t} \tag{4}$$

次に，ラベル列 $\boldsymbol{l}$ の確率は以下で計算できる．

$$p(\boldsymbol{l}|\mathbf{X}, \theta) = \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\boldsymbol{l})} p(\boldsymbol{\pi}|\mathbf{X}, \theta) \tag{5}$$

CTC における前向きアルゴリズムを定義するにあたって，新たな正解ラベル列 $\boldsymbol{l}'$ を導入する．これは $\boldsymbol{l}$ の各ラベル間と最初と最後に blank である 0 を挿入した系列である．従って，$\boldsymbol{l}'$ の長さ $\hat{L}'$ は以下で計算できる．

$$\hat{L}' = 2\hat{L} + 1 \tag{6}$$

$\boldsymbol{l}'$ における s 番目のラベルを l'_s と表すとき，$\boldsymbol{l}$ において対応するラベルは $l_{\lfloor s/2 \rfloor}$ である．ここで，パス $\boldsymbol{\pi}$ の t 番目のラベルと正解ラベル列 $\boldsymbol{l}'$ の s 番目のラベルが対応する全てのパス $\boldsymbol{\pi}$ の確率の総和を，前向き確率 $\alpha_t(s)$ という．$\alpha_t(s)$ は次式のように定義できる．

$$\alpha_t(s) := \sum_{\substack{\boldsymbol{\pi}_{1:t} \in \{0,1,\ldots,2^K\}^t \\ \text{s.t. } \mathcal{B}(\boldsymbol{\pi}_{1:t}) = \boldsymbol{l}_{1:\lfloor s/2 \rfloor}}} \prod_{t'=1}^{t} \phi_{\pi_{t'}, t'} \tag{7}$$

ここで，$\boldsymbol{a}_{1:t}$ は系列 $\boldsymbol{a}$ における最初の t 個の要素を表す．前向き確率は以下の前向きアルゴリズムで計算できる．ま

ず，初期値を次にように定める．

$$\alpha_1(1) = \phi_{0,1} \tag{8}$$
$$\alpha_1(2) = \phi_{l_1,1} \tag{9}$$
$$\alpha_1(s) = 0, \forall s > 2 \tag{10}$$

次に，$t \geq 2$ に対しては，以下の更新式を用いる．

$$\alpha_t(s) = \begin{cases} \text{if } l'_s = \text{blank or } l'_s = l'_{s-2} \\ \quad (\alpha_{t-1}(s) + \alpha_{t-1}(s-1))\, \phi_{l'_s,t} \\ \text{else} \\ \quad (\alpha_{t-1}(s) + \alpha_{t-1}(s-1) + \alpha_{t-1}(s-2))\, \phi_{l'_s,t} \end{cases} \tag{11}$$

式 (8) から (11) を用いて $\alpha_T(s)$ を計算により求めることができる．最後にパスは T フレーム目で L' 番目か $L'-1$ 番目のテイタムに到達している必要がある（それぞれ最後の blank ラベルと最後の blank でないラベルを表す）．以上より正解ラベルの確率 $p(\boldsymbol{l}|\mathbf{X}, \theta)$ は以下のように計算できる．

$$p(\boldsymbol{l}|\mathbf{X}, \theta) = \alpha_T(L'-1) + \alpha_T(L') \tag{12}$$

以上が従来の CTC における確率の計算法である．

3.4 定テンポ制約付き CTC に基づく損失関数

DNN からの出力パス $\boldsymbol{\pi}$ は冗長なラベル列であり，$\mathcal{B}(\boldsymbol{\pi})$ に変換された各ベルは，$\boldsymbol{\pi}$ においてある程度のフレーム数を持っている．本稿では，このフレーム数のことを継続時間と呼ぶ．

従来の CTC に対し，各ラベルの継続時間の集合 $\mathbf{D} = \{d_n\}_{n=1}^{L}$ を導入するとき，最大化する確率は次のようになる．

$$p(\boldsymbol{l}|\mathbf{X}, \theta) = \sum_{\mathbf{D}} p(\boldsymbol{l}, \mathbf{D}|\mathbf{X}, \theta) \tag{13}$$
$$= \sum_{\mathbf{D}} p(\boldsymbol{l}|\mathbf{X}, \theta, \mathbf{D}) p(\mathbf{D}) \tag{14}$$
$$p(\mathbf{D}) = p(d_1) \prod_{n=2}^{L} p(d_n|d_{n-1}) \tag{15}$$
$$p(d_1) = \frac{1}{D_{\max} - D_{\min} + 1} \tag{16}$$
$$p(d_n|d_{n-1}) \propto \exp\left(-\lambda \left| \frac{d_n}{d_{n-1}} - 1 \right| \right) \tag{17}$$
$$d_n \in \{D_{\min}, D_{\min}+1, \ldots, D_{\max}\} \tag{18}$$

ただし，d_n は $\mathcal{B}(\boldsymbol{\pi})$ における n 番目のラベルの継続時間を表し，その範囲は $D_{\min}$ から $D_{\max}$ とする．本稿ではフレーム長が 10 [ms] であるので，BPM は $1500/d_n$ に対応する．また，式 (14) では，$\mathbf{D}$ が $\mathbf{X}$ 及び θ と独立であることを仮定し，式 (17) では，文献 [6] の局所テンポの遷移確率を参考にした．

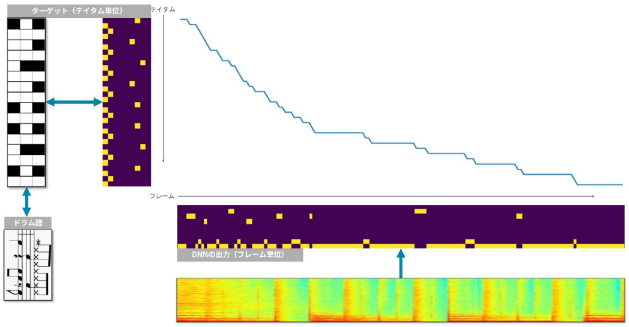

図 3 CTC によるフレーム単位での出力とテイタム単位での出力のアラインメントの例．

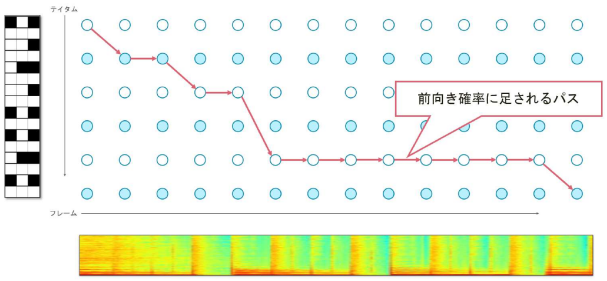

図 4 前向き確率に足し合わされるパスのうちテンポの一貫性が考慮されていない例．

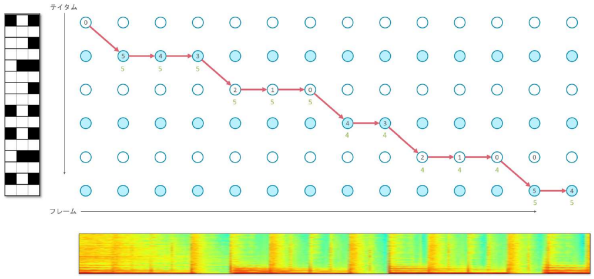

図 5 前向き確率に足し合わされるパスのうちテンポの一貫性が考慮されている例．

提案する CTC における前向き確率は，s 番目のラベルの継続時間 d を変数に含む必要がある．前向き確率はフレーム毎に定義するため，t 番目のフレームが s 番目のラベルの中で，何フレーム目かという情報が必要となる．従って s 番目のラベルが継続中に d から 1 フレーム毎にデクリメントされる変数 c を導入し，本稿では以下カウンタと呼ぶ．以上より前向き確率は以下で表せる．

$$\alpha_t(s, d, c) := \sum_{\substack{\boldsymbol{\pi} \in \mathbb{Z}^T \, s.t. \\ \mathcal{B}(\boldsymbol{\pi}_{1:t}) = \boldsymbol{l}_{1:\lfloor s/2 \rfloor}}} \prod_{t'=1}^{t} \phi_{\pi'_t, t'} \tag{19}$$

なお，提案手法においては継続時間毎のラベル遷移を考えるため，blank ラベル導入しなくても，適切に確率を計算することが可能である．しかし本稿では，blank ラベルを含めた手法を提案する．なぜなら，提案 CTC においても blank ラベルを導入すると，DNN から出力されたパスに対し $\mathcal{B}$ を適用するだけで最終的な採譜システムの出力になり，更に収束速度も向上する．

CTC における前向きアルゴリズムを次にように定める．

$$\alpha_t(s,d,c) = \begin{cases} \text{if } d = c \\ \quad \text{if } l'_s = \text{blank} \\ \quad\quad 0 \\ \quad \text{else if } l'_s = l'_{s-2} \\ \quad\quad \sum_{d'} \alpha_{t-1}(s-1,d',1) \cdot \phi_{l'_s,t} \cdot p(d|d') \\ \quad \text{else} \\ \quad\quad \sum_{d'} \bar{\alpha}_t(s,d,c) \cdot \phi_{l'_s,t} \cdot p(d|d') \\ \text{else if } d > c \\ \quad \text{if } l'_s = \text{blank} \\ \quad\quad \bar{\bar{\alpha}}_t(s,d,c) \cdot \phi_{l'_s,t} \\ \quad \text{else} \\ \quad\quad \alpha_{t-1}(s,d',c+1) \cdot \phi_{l'_s,t} \\ \text{else} \\ \quad 0 \end{cases} \tag{20}$$

$$\bar{\alpha}_t(s,d,c) = \alpha_{t-1}(s-1,d',1) + \alpha_{t-1}(s-2,d',1) \tag{21}$$

$$\bar{\bar{\alpha}}_t(s,d,c) = \alpha_{t-1}(s,d',c+1) + \alpha_{t-1}(s-1,d',c+1) \tag{22}$$

ここで，初期値は次にように定める．

$$\alpha_1(1,d,c) = \phi_{0,1}, \forall d,c \tag{23}$$

$$\alpha_1(2,d,c) = \phi_{l_1,1}, \forall d,c \tag{24}$$

$$\alpha_1(s,d,c) = 0, \forall s > 2 \tag{25}$$

最後にパスは式（12）と同様に，T フレーム目で L' 番目か $L'-1$ 番目のテイタムに到達している必要がある．以上より正解ラベルの確率 $p(\boldsymbol{l}|\mathbf{X},\theta)$ は次式で計算できる．

$$p(\boldsymbol{l}|\mathbf{X},\theta) = \sum_{s=L'-1}^{L'} \sum_{d=D_{\min}}^{D_{\max}} \sum_{c=1}^{d} \alpha_T(s,d,c) \tag{26}$$

継続時間の導入により，図 4 のように傾きが一定でなくテンポの一貫性が考慮されていないパスの確率を足し合わせる時の重みが小さくなる．逆に，図 5 のように傾きが一定でテンポの一貫性が考慮されているパスの確率を足し合わせる時の重みが大きくなるため，定テンポ制約のある出力をしやすいようにネットワークの学習を誘導する．

提案する前向きアルゴリズムでは，$D := D_{\max} - D_{\min} + 1$ とするとき，計算量は $O(TLD^2)$ である．従来の CTC の $O(TL)$ に比べて $O(D^2)$ 倍であるため，多くのテンポに対応させようとしたとき，この計算量は非常に大きくなる．

4. 評価実験

本章では提案法の動作検証結果について報告する．

4.1 実験条件

実験に用いたデータ・ネットワーク・評価尺度について説明する．

4.1.1 データ

提案法の基本的な動作を検証するため，1 秒間の合成ドラム音源（BPM240 で 1 小節相当）を 1 つ使用した．素朴な実装では，計算時間の点でこれ以上の長さの音響信号を扱うことは困難であり，今後の課題とする．

本実験では，計算量の問題で対象とするテンポ幅を制限し，対象となる楽曲のテンポを BPM で表すとき，テイタムの継続時間は $D_{\text{sub}} = 1500/\text{BPM}$ となるため，前後に 2 フレーム程度の幅を持たせて $D_{\min} = \lfloor D_{\text{sub}} - 1 \rfloor$, $D_{\min} = \lfloor D_{\text{sub}} + 2 \rfloor$ と設定した．

4.1.2 ネットワーク

本研究では，潜在特徴を抽出する CNN と双方向長・短期記憶ネットワーク (bidirectional long-short term memory; BLSTM) を用いて DNN を構成した．CNN の各パラメータは，カーネルサイズを 3×3，パディングサイズを 1×1，ストライドを 1 とした．CNN からの出力は 512×4 次元であり，これに 30% の割合の dropout を適用した後, BLSTM に入力される．BLSTM は，3 層かつ 200 次元の隠れ層を持つ．提案モデルの最適化には AdamW [30] を使用し，各パラメータは $\gamma = 0.001$（学習率），$\lambda = 10^{-4}$（weight decay），$\beta_1 = 0.9$，$\beta_2 = 0.98$，$\varepsilon = 10^{-9}$ とした．提案モデルは PyTorch v1.8.1 を用いて実装した．

4.1.3 評価尺度

提案手法に対しては，単純な F 値を用いた評価は適していない．推定テイタム数と正解テイタム数が同じという保証が無いことと，一部分が推定されなかったときにそれ以降が不正解となってしまうことが理由である．従って，本研究では単語誤り率（Word Error Rate; WER）を参考にした以下の誤り率を評価尺度を用いる．

$$Err = 100 \cdot \frac{Ins. + Del. + Sub.}{\hat{L}} \ [\%] \tag{27}$$

ここで，$Ins.$，$Del.$，$Sub.$ はそれぞれ，テイタム単位でのドラムラベルの挿入，削除，置換誤りを表し，$\hat{L}$ は正解テイタム数を表す．

4.2 実験結果

実験結果を図 6 に示す．この実験結果は訓練データに対するものであるが，従来の CTC と比べてドラムが無い部分を推定できていることを確認できる．誤り率はそれぞれ，$Err = 25\%$ と $Err = 0\%$ である．

正解ドラム譜と DNN の出力 $\mathbf{H}$ のアラインメントを図 7 に示す．従来の CTC では，アラインメントの傾きが一定でないため，出力記号の継続時間が一定に保たれていないことが分かる．特に，傾きが急なところでドラムラベルの削除誤りががある．提案する CTC では，アラインメントの傾きが一定に近く，ドラムラベルの削除誤りは存在しない．

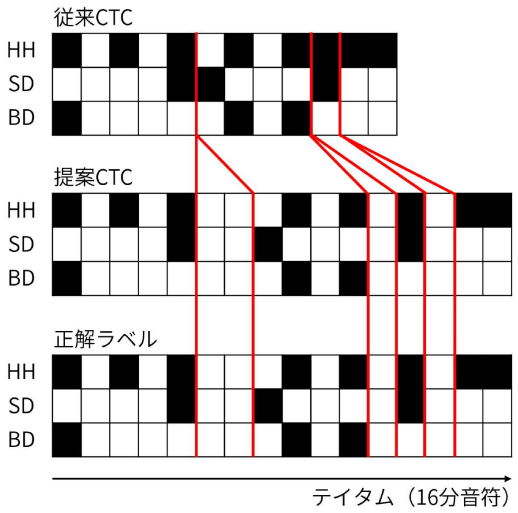

図 6 従来法と提案法による採譜結果の比較：縦軸はドラムの種類を表し，横軸はテイタム単位を表す．黒い部分は，ドラムの打音があることを表す．

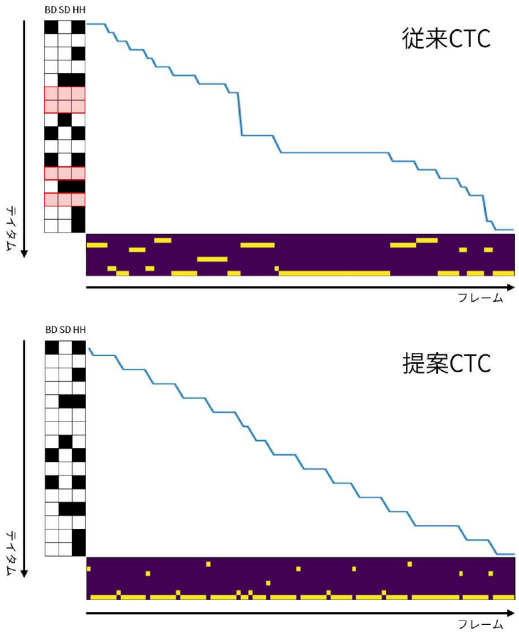

図 7 従来法と提案法におけるラティスの比較：縦軸はテイタム単位の正解ドラム譜を表し，横軸はフレーム単位の $\mathbf{H}$ を表す．推定誤りを赤く囲まれたテイタムで示す．

5. おわりに

本稿では，テンポの一定性を考慮した CTC に基づくテイタム単位での自動ドラム採譜手法を提案した．本研究の主要な貢献は，自動ドラム採譜タスクにおいて，楽曲の音響信号から直接楽譜上のリズムを特定できる形での出力を可能にしたことと，定テンポ制約を導入した CTC を提案してドラム打音が存在しない部分の推定を可能にしたことである．提案法では，CTC において前向き確率の潜在変数に出力記号の継続時間を導入した新しい損失関数に基づいており，継続時間系列が従う分布 (15) と継続時間の遷移確率 (17) に変更を加えることで，自動採譜タスク以外にも応用可能である．例えば，手書き文字認識のための CTC においては，1 文字当たりの画像の横幅がテンポに対応するため，文字の大きさの一定性の制約を導入する目的での利用が考えられる [31]．一方，潜在変数の増加に伴い計算量が非常に大きくなるため，大きなデータセットによるネットワークの学習を行うための計算の高速化の必要性も明らかになった．

今後は，前向きアルゴリズムの見直しや計算パスを削減により高速化を図り，大きなデータセットでの評価を行う．提案法が現実的な速度で実行可能になれば，対応するペアデータの作成が簡単であることから，従来より大幅に大きいデータを用いることができ，より精度の良いモデルを作成できると期待できる．また，学習時だけでなく推論時にも継続時間を考慮するため，式 (2) に Viterbi アルゴリズムを導入する．さらに，テイタムを同時に推定するモデルなどとの比較により，ドラム採譜において最適なモデルについて検討する．

謝辞 本研究の一部は，JST PRESTO No. JPMJPR20CB 及び科研費 No. 19H04137, 21H03572, 21K02846, 21K12187, 22H03661 の支援を受けた．

参考文献

[1] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore and D. Eck: Onsets and Frames: Dual-Objective Piano Transcription, *Proc. International Society for Music Information Retrieval (ISMIR)*, pp. 50–57 (2018).

[2] R. Stables, J. Hockman and C. Southall: Automatic Drum Transcription using Bi-directional Recurrent Neural Networks., dblp (2016).

[3] R. Vogl, M. Dorfer, G. Widmer and P. Knees: Drum Transcription via Joint Beat And Drum Modeling using Convolutional Recurrent Neural Networks, *Proc. International Society for Music Information Retrieval (ISMIR)*.

[4] E. Nakamura, E. Benetos, K. Yoshii and S. Dixon: Towards Complete Polyphonic Music Transcription: Integrating Multi-Pitch Detection and Rhythm Quantization, *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, pp. 101–105 (2018).

[5] K. Shibata, E. Nakamura and K. Yoshii: Non-local musical statistics as guides for audio-to-score piano transcription, *Information Sciences*, Vol. 566, pp. 262–280 (2021).

[6] F. Krebs, S. Böck and G. Widmer: An Efficient State-Space Model for Joint Tempo and Meter Tracking., *Proc. International Society for Music Information Retrieval (ISMIR)*, pp. 72–78 (2015).

[7] S. Böck and M. E. Davies: Deconstruct, Analysis, Reconstruct: How To Improve Tempo, Beat, and Downbeat Estimation, *Proc. International Society for Music Information Retrieval (ISMIR)*.

[8] R. Ishizuka, R. Nishikimi, E. Nakamura and K. Yoshii: Tatum-Level Drum Transcription Based on a Convolu-

tional Recurrent Neural Network with Language Model-Based Regularized Training, *Asia-Pacific Signal and Information Processing*.

[9] R. G. C. Carvalho and P. Smaragdis: Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 151–155 (2017).

[10] M. A. Román, A. Pertusa and J. Calvo-Zaragoza: An End-to-end Framework for Audio-to-Score Music Transcription on Monophonic Excerpts., *Proc. International Society for Music Information Retrieval (ISMIR)*, pp. 34–41 (2018).

[11] R. Nishikimi, E. Nakamura, M. Goto and K. Yoshii: End-to-end melody note transcription based on a beat-synchronous attention mechanism, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, pp. 26–30 (2019).

[12] L. Liu, V. Morfi and E. Benetos: Joint multi-pitch detection and score transcription for polyphonic piano music, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 281–285 (2021).

[13] A. Graves, S. Fernandez, F. Gomez and J. Schmidhuber: Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks, *ICML*, pp. 369–376 (2006).

[14] S. Kim, T. Hori and S. Watanabe: Joint CTC-attention based end-to-end speech recognition using multi-task learning, *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 4835–4839 (2017).

[15] 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).

[16] C. Jacques and A. Roebel: Automatic drum transcription with convolutional neural networks, *21th International Conference on Digital Audio Effects, Sep 2018, Aveiro, Portugal* (2018).

[17] C. Jacques and A. Roebel: Data augmentation for drum transcription with convolutional neural networks, *IEEE European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (2019).

[18] 上田舜，柴田健太郎，和田雄介，錦見亮，中村栄太，吉井和佳：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，研究報告エンタテインメントコンピューティング (EC)， Vol. 2019, No. 26, pp. 1–6 (2019).

[19] Y. Wang, J. Salamon, M. Cartwright, N. J. Bryan and J. P. Bello: Few-shot drum transcription in polyphonic music, *arXiv preprint arXiv:2008.02791* (2020).

[20] R. Ishizuka, R. Nishikimi and K. Yoshii: Global Structure-Aware Drum Transcription Based on Self-Attention Mechanisms, *Signals*, Vol. 2, No. 3, pp. 508–526 (2021).

[21] C. Lea, R. Vidal, A. Reiter and G. D. Hager: Temporal convolutional networks: A unified approach to action segmentation, *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, Springer, pp. 47–54 (2016).

[22] M. E. Daveis and S. Böck: Temporal convolutional networks for musical audio beat tracking, *IEEE European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (2019).

[23] 鎌倉大地，大山偉永，吉井和佳：マルチタスク学習に基づくドラム採譜と拍節構造推定，第 84 回全国大会講演論文集， Vol. 2022, No. 1, pp. 517–518 (2022).

[24] M. Cartwright and J. P. Bello: Increasing drum transcription vocabulary using data synthesis, *Proc. International Conference on Digital Audio Effects (DAFx)*, pp. 72–79 (2018).

[25] K. Choi and K. Cho: Deep unsupervised drum transcription, *arXiv preprint arXiv:1906.03697* (2019).

[26] R. G. C. Carvalho and P. Smaragdis: Towards End-to-End Polyphonic Music Transcription: Transforming Music Audio Directly to a Score, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 151–155 (2017).

[27] M. A. Román, A. Pertusa and J. Calvo-Zaragoza: An End-To-End Framework for Audio-To-Score Music Transcription on Monophonic Excerpts, *Proc. International Society for Music Information Retrieval (ISMIR)*, pp. 34–41 (2018).

[28] R. Nishikimi, E. Nakamura, S. Fukayama, M. Goto and K. Yoshii: Automatic Singing Transcription Based on Encoder-Decoder Recurrent Neural Networks with a Weakly-Supervised Attention Mechanism, *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, pp. 161–165 (2019).

[29] T. Zhao: Viterbi Accelerated Training for CTC Series Topologies.

[30] I. Loshchilov and F. Hutter: Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017).

[31] M. Ibrayim, W. Simayi and A. Hamdulla: Unconstrained online handwritten Uyghur word recognition based on recurrent neural networks and connectionist temporal classification, *International Journal of Biometrics*, Vol. 13, No. 1, pp. 51–63 (2021).