Vol.2019-MUS-124 No.8
2019/8/28

# ビート同期注意機構に基づく歌声のリズム採譜

錦見 亮[1,a)]　中村 栄太[1,b)]　吉井 和佳[1,c)]

**概要**：本稿では，ポピュラー音楽を対象に，時間的に量子化されていない楽曲と歌声のピアノロールを入力として，歌声の音符（音高・音価）系列を推定する方法について述べる．我々は，歌声採譜を機械翻訳や音声認識と同様の系列変換問題として捉え，近年大きな成果を上げている注意機構を用いたエンコーダ・デコーダモデルの適用を試みる．しかし， エンコーダは瞬時的特徴の抽出は得意だが，経時的特徴の抽出は不得意であるため，単純に，音符単位で出力を行うデコーダを用いることは得策ではない．また，小節線やビート位置などの拍節構造を推定することができない．系列変換問題ではあるが，時間的な情報が本質的な役割を果たす自動採譜特有の問題を解決するために，我々は，テイタム単位で音符ラベル系列と拍節構造ラベル系列を出力するデコーダを提案する．さらに，注意機構を通じた入力系列と出力系列の対応付けの際に，注意重みの重心が単調増加かつ一定間隔に並ぶよう制約をかけたビート同期注意機構を提案する．実験では，提案するリズム採譜手法の基本的な能力を検証した．

## 1. はじめに

歌声採譜は，音楽音響信号を入力とし，歌声の楽譜を推定するタスクである．メロディを担う歌声は楽曲中で最も主要な構成要素であり，楽曲の印象に密接に関連しているため，歌声解析は音楽情報処理の基盤技術となっている．歌声採譜が実現すれば，人間や計算機にとって扱いが容易な離散記号でのみで記述された楽譜形式を取得することができる．また，推定された楽譜は，歌声生成や音楽文法解析，ハミング検索や能動的音楽鑑賞システムなど様々な場面で応用できる．

歌声採譜を行うには，ピアノロール推定とリズム採譜を組み合わせる方法が考えられる．ピアノロール推定では，楽曲や歌声の音響信号から連続的な F0 軌跡を推定し，各フレームごとに F0 を半音単位に量子化するだけのものもあるが，音高・オンセット時刻・継続時間の三つ組で表現される音符イベントを推定することが望ましい．ただし，この段階では，音符イベントの音高は離散的であるが，オンセット時刻および継続時間は秒単位で表現されている．完全な楽譜を推定するには，リズム採譜において，音符イベントを時間的に量子化して楽譜上における音符のオンセット位置や音価を推定する必要がある．同時に，楽譜上でのオンセット位置や音価を表す基本単位となる拍子やビートなどの拍節構造も推定する必要がある．

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

歌声採譜は，楽曲のスペクトル系列から音符（音高と音価）系列への系列変換問題であるので，同型の問題である機械翻訳 [1, 2] や音声認識 [3–5] で使用されるアプローチが有効であると考えられる．特に，注意機構に基づくエンコーダ・デコーダモデルによる End-to-End アプローチが有望である．エンコーダは，入力系列を同じ長さの潜在表現ベクトル系列に変換する．デコーダは，潜在表現ベクトル系列を受け取り，終端記号が現れるまで出力記号を再帰的に予測する．このとき，各ステップにおいて，潜在表現ベクトル系列との関連度（注意重み）を計算し，潜在表現ベクトル系列の重みづけ平均を取ることにより，適切な位置の潜在表現ベクトルを出力記号に変換することができる．双方ともに，通常は，再帰型ニューラルネットワーク (recurrent neural network: RNN) を用いて実装される．

我々は以前，楽曲の音響信号から歌声があらかじめ分離されているという仮定のもとで，標準的な注意機構を持つエンコーダ・デコーダモデルを用いて歌声のスペクトル系列から音符系列を推定する手法 [6] を提案した．しかし，エンコーダは，音高のような瞬時的特徴の抽出は得意であるが，音価のような経時的特徴の抽出が不得意であるため，音符単位で出力を行うデコーダでは，注意機構による入出力系列のアラインメントが失敗しやすく，正しい音価の推定が特に困難であった．たとえ，エンコーダに長短期記憶 (long short-term memory: LSTM) を用いていても，音価を表現する特徴量を数十から数百ステップに渡って伝搬させることは難しいと考えられる．この問題に対し，学習データの一部だけでも，音響信号中での各音符のオンセッ

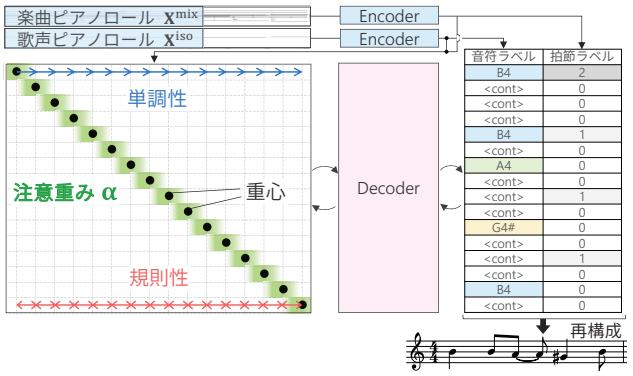


**図 1** 歌声のリズム採譜のためのビート同期注意機構に基づくエンコーダ・デコーダモデル.

ト時刻が分かっている場合に，注意重みに対する損失関数を導入して，適切なアラインメントに誘導する方法を提案した．しかし，そのような教師データを準備するには限界があり，完全な楽譜を得るためには，追加で拍節構造の推定が必要であった．

本稿では，音楽音響信号に対して正確なピアノロール推定（音高推定）ができているという仮定のもとで，歌声と楽曲のピアノロールに対するリズム採譜に取り組み，注意機構を持つエンコーダ・デコーダモデルが，出力記号の経時的な情報（音価）を推定する能力を持ちえるかを検証する．これは，機械翻訳や音声認識などには存在しない，自動採譜特有の興味深い疑問である．提案するリズム採譜モデルの肝は，音符単位ではなく，テイタム（16 分音符）単位のデコーダを用いる点である（図 1）．まず，二つのエンコーダが，歌声と楽曲のピアノロールをそれぞれ入力として受け取り，潜在表現ベクトル系列に変換する．次に，共通のデコーダは，二つの潜在表現ベクトル系列を参照しながら，テイタム単位の楽譜情報系列（音符ラベル系列と拍節構造ラベル）を再帰的に出力する．デコーダから音価を直接出力することをやめ，デコーダの出力を比較的局所的な情報に限定する代わりに，注意機構を通じた入出力のアラインメントによって音価を推定する．また、この仕組みにより，同時に拍節構造を推定することが可能になる．

さらに，本研究では，注意重みが**単調性**と**規則性**を持つよう誘導する損失関数に基づくビート同期注意機構を提案する．採譜問題において，入出力系列の対応順序は逆転せず，典型的なポピュラー楽曲ではテンポ変動は小さいため，注意重みのピークが立つ場所は昇順（単調性）かつ一定間隔（規則性）に並ぶことが期待される．単調性は音声認識でも成立し，注意重みの計算式 [7] やネットワーク構造 [7] を工夫して実現する方法が提案されている．一方，本研究では，単調性制約と規則性制約の両方を損失関数として実装し，デコーダの出力に対する損失関数と同時に最小化することを試みる．ここで，注意重みに対する損失関数は自己組織的に計算でき，教師データを必要としない．

## 2. 関連研究

本章では，歌声に関するものを中心に自動楽曲採譜の関連研究を述べる．

### 2.1 ピアノロール推定

ピアノロールを推定する試みはこれまでに多く行われてきた [8–13]．歌声のピアノロール推定には，ルールやフィルターに基づくもの [8,9] や隠れマルコフモデル (hidden Markov model: HMM) に基づくもの [10,11] があり，事前に推定された歌声 F0 軌跡を対象にすることが多い．これは，ビブラートやグリッサンド等の歌唱表現により歌声の音高やオンセットは変動が大きく，直接スペクトログラムからピアノロールを推定するのが難しいためである．一方，ピアノ等の音高が安定しオンセットが明瞭な楽器では，歌声 F0 軌跡を介さずにピアノロールを推定することもできる．例えば，確率的潜在要素解析は [12]，スペクトログラムから直接音高の量子化を行い，HMM に基づくノートトラッキングと組み合わせてピアノロールが推定される．また，最近では DNN を用いて直接ピアノロール推定する手法 [13,14] が高い推定精度を達成している．

### 2.2 楽譜推定

音高・時間方向ともに量子化された楽譜を推定する研究も存在する [6,15–21]．中村らは [17]，ピアノロールから音符の音価とオンセット位置を推定するリズム採譜を行うため，拍節 HMM [15,16] を応用している．錦見ら [18] や中村ら [19] は，歌声 F0 軌跡から音高離散化と音価・オンセット位置推定を行うため，所与のビート時刻に基づく階層セミマルコフモデルを提案している．近年は，DNN を用いて End-to-End で音響信号から楽譜を直接推定する試みも盛んになりつつある．Carvalho ら [20] は，ピアノ音に対して一次元畳み込みニューラルネットワーク (convolutional neural network: CNN) を用いて抽出した特徴量から Lilypond 形式 [22] の記号を推定する Sequence-to-Sequence モデル [23] を提案している．Román ら [21] は，connectionist temporal classification (CTC) [24] を用いて，振幅スペクトログラムから調，音高，音価，拍子などの音楽記号の系列を推定する手法を提案している．錦見ら [6] は，弱教師あり注意機構を用いたエンコーダ・デコーダモデルに基づき歌声から音高・音価系列を推定する手法を提案している．弱教師あり注意機構は，音符のオンセット時刻を用いて注意重みを適切な値に誘導するもので，歌声と音符のアラインメントや学習時間を改善する効果がある．

## 3. 提案手法

本章では，注意機構に基づくエンコーダ・デコーダモデルを用いて歌声のリズム採譜を行う方法を説明する．

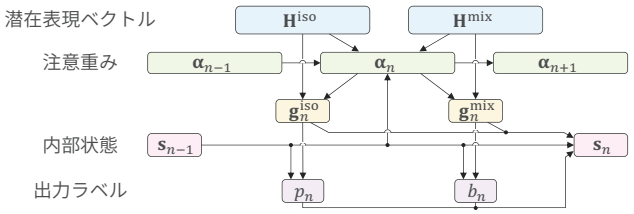


**図 2** 提案法における注意機構付きデコーダのアーキテクチャ．

### 3.1 問題設定

我々は，フレームレベルのピアノロールから，歌声の楽譜を推定する問題に取り組む．$T$ と $N$ をそれぞれ入力ピアノロールの時間フレーム数および出力楽譜の 16 分音符単位のテイタム数とする．提案法の入力は歌声の演奏ピアノロール $\mathbf{X}^{\text{iso}} \in \{0,1\}^{2\times 128\times T}$ と楽曲のコンデンス演奏ピアノロール $\mathbf{X}^{\text{mix}} \in \{0,1\}^{2\times 128\times T}$ である．各ピアノロールは音高とオンセットに対応する二つの行列から構成される．音高行列はノートがある時間・周波数ビンは 1，その他のビンは 0 である．また，オンセット行列はノートのオンセットがある時間・周波数ビンのみが 1 で，その他のビンは 0 である．異なるパートで音高が重複する場合，短い音符のオンセット情報が失われてしまうことを防ぐため，オンセット行列も併せて用意した．

出力は歌声の楽譜を表すテイタム単位のラベル系列 $\mathbf{Y} = [y_1,\dots,y_N] = [(p_1,b_1),\dots,(p_N,b_N)]$ である．各 $p_n \in \{1,\dots,K,\langle\texttt{cont}\rangle,\langle\texttt{sos}\rangle,\langle\texttt{eos}\rangle\}$ は音符系列を表すラベルである．$K$ は出力として想定されている音高の種類数（休符含む）を表し，$p_n = k \in \{1,\dots,K\}$ は $n$ 番目のテイタムが音符のオンセットかつ音高が $k$ であることを表す．本稿では，音高の種類は E2 から G5 までの 40 半音と休符（$K = 41$）と設定する．また，$p_n = \langle\texttt{cont}\rangle$ は $n$ 番目のテイタムにおいて音符が継続していることを表す．一方，各 $b_n \in \{\langle\texttt{downbeat}\rangle,\langle\texttt{beat}\rangle,\langle\texttt{nobeat}\rangle,\langle\texttt{sos}\rangle,\langle\texttt{eos}\rangle\}$ は拍節構造を表すラベルである．$b_n = \langle\texttt{downbeat}\rangle$，$b_n = \langle\texttt{beat}\rangle$，および $b_n = \langle\texttt{nobeat}\rangle$ はそれぞれ，$n$ 番目のテイタムがビート，小節線，およびそのどちらでもないことを表す．また，音符ラベルと拍節構造ラベルの語彙に含まれる $\langle\texttt{sos}\rangle$ と $\langle\texttt{eos}\rangle$ は出力系列の開始と終了を表す特殊ラベルである．

### 3.2 エンコーダ

提案モデルは $\mathbf{X}^{\text{iso}}$ 用と $\mathbf{X}^{\text{mix}}$ 用の二つのエンコーダを持つ．二つのエンコーダは同じアーキテクチャであり，前段の 2 次元 CNN と後段の LSTM で構成される．CNN は $\mathbf{X}^{\text{iso}}$ と $\mathbf{X}^{\text{mix}}$ それぞれに対して，音高行列とオンセット行列を同じサイズの一つの行列に変換する．その行列はバッチ正規化が行われたのち，後段の LSTM に渡され，潜在表現ベクトル系列 $\mathbf{H}^{\text{iso}} = [\mathbf{h}_1^{\text{iso}},\dots,\mathbf{h}_T^{\text{iso}}] \in \mathbb{R}^{E\times T}$ と $\mathbf{H}^{\text{mix}} = [\mathbf{h}_1^{\text{mix}},\dots,\mathbf{h}_T^{\text{mix}}] \in \mathbb{R}^{E\times T}$ へ変換される．ここで，$E$ は潜在表現ベクトルの次元を表す．

### 3.3 注意機構付きデコーダ

デコーダは潜在表現ベクトル系列 $\mathbf{H} = \mathbf{h}_{1:T} \in \mathbb{R}^{2E\times T}$ から出力記号系列 $\mathbf{Y}$ を予測する．ここで，$\mathbf{h}_t$ は潜在表現ベクトル $\mathbf{h}_t^{\text{iso}}$ と $\mathbf{h}_t^{\text{mix}}$ の連結ベクトルである．デコーダは単方向 LSTM で構成し，以下の処理を再帰的に行う（図 2）．

$$\boldsymbol{\alpha}_n = \text{Attend}(\mathbf{s}_{n-1},\boldsymbol{\alpha}_{n-1},\mathbf{H}) \tag{1}$$

$$\mathbf{g}_n = \sum_{t=1}^{T} \alpha_{nt}\mathbf{h}_t \tag{2}$$

$$y_n = \text{Generate}(\mathbf{s}_{n-1},\mathbf{g}_n) \tag{3}$$

$$\mathbf{s}_n = \text{Recurrency}(\mathbf{s}_{n-1},\mathbf{g}_n,y_n) \tag{4}$$

ここで，$\boldsymbol{\alpha}_n \in \mathbb{R}^T$ は注意重みベクトル，$\mathbf{s}_n \in \mathbb{R}^D$ は $n$ 番目のタイムステップにおけるデコーダ LSTM の内部状態を表す．また，ベクトルや行列に対して演算を行う関数 Attend，Generate，Recurrency については以下に説明する．

式 (1) と (2) は注意機構における演算を表す．注意重み $\boldsymbol{\alpha}_n \in \mathbb{R}^T$ は合計が 1 になるよう正規化された確率ベクトルであり，内部状態 $\mathbf{s}_n$ に対する潜在表現ベクトル系列 $\mathbf{H}$ の関連度合を表す．$\boldsymbol{\alpha}_n$ の各要素は以下のように計算される．

$$\alpha_{nt} = \frac{\exp(e_{nt})}{\sum_{t'=1}^{T}\exp(e_{nt'})} \tag{5}$$

$$e_{nt} = \text{Score}(\mathbf{s}_{n-1},\mathbf{h}_t,\boldsymbol{\alpha}_{n-1}) \tag{6}$$

ここで，Score は正規化前の注意重みを計算する関数である．音符と拍節構造の両方にとって重要な情報が含まれるように $\mathbf{H}^{\text{iso}}$ と $\mathbf{H}^{\text{mix}}$ の両方から一つの共有注意重み行列 $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{1:N} \in \mathbb{R}^{N\times T}$ が計算される．Score 関数として，次式で計算される畳み込み演算を行うものを用いた [3]．

$$\mathbf{f}_n = \mathbf{F} * \boldsymbol{\alpha}_{n-1} \tag{7}$$

$$e_{nt} = \mathbf{w}^\top \tanh\left(\mathbf{W}\mathbf{s}_{n-1} + \mathbf{V}\mathbf{h}_t + \mathbf{U}\mathbf{f}_{nt} + \mathbf{b}\right) \tag{8}$$

ここで，“$*$” は 1 次元畳み込み演算，$\mathbf{F} \in \mathbb{R}^{C\times I}$ は畳み込みカーネルの集合，$\mathbf{f}_n \in \mathbb{R}^{C\times T}$ は畳み込み演算の結果，$C$ と $I$ はカーネル数とカーネルサイズを表す．$\mathbf{w} \in \mathbb{R}^A$ は重みベクトル，$\mathbf{W} \in \mathbb{R}^{A\times D}$，$\mathbf{V} \in \mathbb{R}^{A\times 2E}$，$\mathbf{U} \in \mathbb{R}^{A\times I}$ は重み行列，$\mathbf{b} \in \mathbb{R}^A$ はバイアスベクトル，$A$ は $\mathbf{W}$，$\mathbf{V}$，$\mathbf{U}$ の行数および $\mathbf{b}$ の要素数を表す．

式 (2) によって，$\mathbf{h}^{\text{iso}}$ と $\mathbf{h}^{\text{mix}}$ から計算される重み付き和 $\mathbf{g}_n^{\text{iso}} = [g_{n1},\dots,g_{nE}]^\top$ と $\mathbf{g}_n^{\text{mix}} = [g_{n,E+1},\dots,g_{n,2E}]^\top$ の結合ベクトルを $\mathbf{g}_n \in \mathbb{R}^{2E}$ とすると，式 (3) で表される $y_n$ の生成過程は以下のとおりである．

$$\boldsymbol{\phi}^{(n)} = \text{softmax}\left(\mathbf{P}^{\text{p}}\mathbf{s}_{n-1} + \mathbf{Q}^{\text{p}}\mathbf{g}_n^{\text{iso}} + \mathbf{b}^{\text{p}}\right) \tag{9}$$

$$p_n = \underset{p\in V^{\text{p}}}{\text{argmax}}\left(\phi_p^{(n)}\right) \tag{10}$$

$$\boldsymbol{\psi}^{(n)} = \text{softmax}\left(\mathbf{P}^{\text{b}}\mathbf{s}_{n-1} + \mathbf{Q}^{\text{b}}\mathbf{g}_n^{\text{mix}} + \mathbf{b}^{\text{b}}\right) \tag{11}$$

$$b_n = \underset{b\in V^{\text{b}}}{\text{argmax}}\left(\psi_b^{(n)}\right) \tag{12}$$

ここで，$\mathbf{P}^{\star} \in \mathbb{R}^{|V^{\star}| \times D}$，$\mathbf{Q}^{\star} \in \mathbb{R}^{|V^{\star}| \times E}$ は重み行列，$b^{\star} \in \mathbb{R}^{|V^{\star}|}$ はバイアスベクトルである．また，“$\star$” は “p” か “b” のいずれかである．$\mathbf{g}_n^{\mathrm{iso}}$ は歌声の音符ラベル系列を推定するために用いられ，$\mathbf{g}_n^{\mathrm{mix}}$ は拍節構造ラベル系列を推定するために用いられる．

式 (4) は状態 $\mathbf{s}_n$ の再帰的な計算を表す．モデル学習時には teacher forcing を採用する．つまり，正解の音符ラベルと拍節構造ラベルを別々に one-hot ベクトルに変換したのち，これらの結合ベクトルを $y_n$ として用いる．提案モデルは，各 $y_n$ に対するクロスエントロピーと次節で説明するロスの合計を最小化することで最適化される．推論時には，式 (10) と (12) で得られた 1 ステップ前での音符ラベルと拍節構造ラベルを one-hot ベクトルに変換したのち，これらの結合ベクトルを現在の音符ラベルと拍節構造ラベルの予測に用いる．この再帰的な計算は，出力系列長が予め設定した最大長に到達するか，音高ラベルまたは拍節構造ラベルの予測結果が ⟨eos⟩ である場合に終了する．

### 3.4 注意重みに対するロス関数

注意重みが単調性と規則性の制約を満たすように誘導する損失関数を導入する．注意機構において入出力のアライメントが適切に行われている場合，注意重み $\boldsymbol{\alpha}_n$ は時間軸上で局所的に集中する傾向がある．そこで各 $\boldsymbol{\alpha}_n$ の代表点として，次式で定義される重心を採用する．

$$c_n = \sum_{t=1}^{T} t \cdot \alpha_{nt} \tag{13}$$

注意重みに対しての単調性制約を課す損失関数は次式で与えられる．

$$\mathcal{L}^{\mathrm{mono}} = \frac{1}{N-1} \sum_{n=1}^{N-1} \mathrm{ReLU}\left(-\Delta c_n\right) \tag{14}$$

ここで，$\Delta c_n = c_{n+1} - c_n$ は隣り合う重心の差である．ReLU は正規化線形関数 $\mathrm{ReLU}\left(x\right) = \max\left(0, x\right)$ である．式 (14) によって，隣り合う重心位置が逆転する場合にのみコストを課す．一方，注意重みに対して規則性制約を課す損失関数は次式で与えられる．

$$\mathcal{L}^{\mathrm{reg}} = \frac{1}{N-2} \sum_{n=1}^{N-2} \left|\Delta c_{n+1} - \Delta c_n\right|^2 \tag{15}$$

式 (15) は重心間隔の二乗誤差をコストとして課すことで，重心位置が時間方向に一定間隔で並ぶよう誘導する．

## 4. 評価実験

本章では提案法に対する精度評価の結果を報告する．

### 4.1 実験データ

提案法を評価するため，RWC Music Database [25] に含まれるアノテーションが正確であるポピュラー音楽 54 曲を使用した．入力ピアノロールと出力楽譜はそれぞれアノテーションデータ [26] 内の同期 MIDI と時間離散化済み MIDI を用いて作成した．1 曲全体を提案法に入力するのは，時間・空間計算量の面で難しいので，窓幅 8 秒，シフト幅 1 秒で入力データを区間に分割した．また，正解楽譜も各区間に対応するように，アノテーションデータ内のビート時刻を使って分割した．入力データ区間の終端をまたぐようなテイタムは正解楽譜から除外した．休符しか含まれない区間は使用しなかった．

### 4.2 実験設定

各エンコーダの CNN の各パラメータは, カーネルサイズを $1 \times 5$，時間方向のパディングサイズを 1，ストライドを 1 とした．また，バッチ正規化の各パラメータは，$\epsilon = 1e^{-5}$ およびモメンタムを 0.1 とした．各エンコーダの LSTM は 3 層かつ $100 \times 2$ 次元の隠れ層を持ち，いずれの層も 20% の割合で dropout を適用した．デコーダは 100 次元の隠れ層を持つ 1 層の単方向 LSTM で構成した．注意機構における畳込み演算のパディングサイズとストライドはそれぞれ 50 と 1 に設定した．また，その他のパラメータは $C = 10$，$I = A = 100$ とした．提案モデルのパラメータ最適化には Adam [27] を使用した．Adam の各パラメータは $\alpha = 0.001$（学習率），$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\varepsilon = 10^{-8}$ とした．また，係数 $10^{-5}$ の荷重減数（L2 正則化）と閾値 5.0 の gradient clipping を行った．全結合層の重みパラメータは一様分布 $\mathcal{U}(-0.1, -0.1)$ からのサンプルで初期化し，バイアスパラメータはゼロ初期化した．CNN のカーネルやエンコーダ・デコーダの重みパラメータは He らの手法 [28] で初期化した．バッチサイズは 100，エポック数は 100 とした．提案モデルは PyTorch v1.1.0 を用いて実装した．

### 4.3 評価尺度

音符単位での評価を行うために，[17] で提案された評価法を用いた．この評価法は，音高誤り率 $E_{\mathrm{p}}$，削除誤り率 $E_{\mathrm{m}}$，挿入誤り率 $E_{\mathrm{e}}$，オンセット誤り率 $E_{\mathrm{on}}$，オフセット誤り率 $E_{\mathrm{off}}$，およびこれら 5 つの値の平均値 $E_{\mathrm{mean}}$ が計算される．ただし，この評価法では休符および拍節構造は考慮されていない．よって，提案法が推定したラベル系列から楽譜を構成する際に拍節構造ラベルは無視した．また，正解楽譜や推定ラベルが ⟨cont⟩ から始まる場合，その ⟨cont⟩ に対応する音高が不明なので休符の継続を表すものとした．

### 4.4 実験結果

表 1 に注意重み損失の有無で比較した提案法の採譜精度示す．単調性制約を注意重みに課すと，全ての評価基準において精度が向上しており，この制約が有効であること

表 1　音符推定誤り率 [%].

| 注意重み損失 | | 評価値 [17] | | | | | |
|---|---|---|---|---|---|---|---|
| 単調性 | 規則性 | $E_{\mathrm{p}}$ | $E_{\mathrm{m}}$ | $E_{\mathrm{e}}$ | $E_{\mathrm{on}}$ | $E_{\mathrm{off}}$ | $E_{\mathrm{mean}}$ |
| | | 0.17 | 2.69 | 5.37 | 15.52 | 8.80 | 6.51 |
| ✓ | | 0.11 | 1.53 | 3.88 | 14.79 | 7.66 | 5.59 |
| | ✓ | 0.23 | 3.73 | 4.89 | 27.28 | 11.61 | 9.55 |

が分かった．一方で，規則性制約を課すと精度が悪くなった．これは，制約が強すぎるためにモデルが適切な注意重み行列を探索する妨げになっているとこが原因だと考えられる．また，音高誤り率が他の評価値に比べて非常に低いが，これは入力に歌声のピアノロールを与えているため，音高推定がオンセット推定や拍節構造推定に比べてに簡単であることに起因する．

図 3 に提案法によって推定された楽譜と注意重みの例を示す．注意重み損失を導入せずに楽譜を推定すると，注意重みの重心順序が逆転し，逆転箇所で音価を分割する誤りが生じていることから，単調性制約が有効であることが分かる．一方，規則性制約を導入すると，制約が強すぎて学習が進まず注意重みが曖昧になっている．これに対して，他の推定結果では音符のオンセット周辺で注意重みのピークが立っている．このことから，音価推定には注意重みのピークが重要であり，注意重みが曖昧な規則性制約ありの楽譜では音価誤りが生じている．また，規則性制約ありの楽譜で見られるような小節線の位相ずれは，注意重み損失の有無にかかわらず他のテストデータに対しても多く見られた．よって，小節線の推定精度を向上させるためには，ピアノロールの代わりに音響信号を入力し，ドラムなどの打楽器音を参照すべきてあると考えられる．

## 5. おわりに

本稿では，ビート同期注意機構に基づくエンコーダ・デコーダモデルを用いて，時間的に量子化されていない歌声と楽曲のピアノロールから歌声の楽譜を推定する方法について述べた．提案法では，音符と拍節構造の同時推定を行うため，テイタム単位で楽譜情報を再帰的に出力するデコーダを設計した．さらに，フレーム単位の入力系列とテイタム単位の出力系列とのアラインメントが適切に学習されるよう，注意重みの単調性および規則性に関する損失関数を提案した．実験により，提案するエンコーダ・デコーダモデルが，経時的な情報である音価を推定する能力を有すること，単調性に関する損失関数が推定精度向上に効果的であることが分かった．一方，規則性に関する損失関数は，我々の期待に反して，推定精度を悪化させる傾向があり，さらなる調査が必要である．

今後，入力を正確なピアノロールから歌声や楽曲の音響信号のスペクトル系列に変更し，End-to-End で歌声採譜ができるかを検証する予定である．これにより，歌声の楽譜推定精度が，前処理のピアノロール推定精度に影響されなくなる．もし，歌声分離モデルを統合できれば，市販楽曲をそのまま入力するだけで歌声の楽譜を得ることが可能になる．発展的な内容として，音符や拍節構造だけではなくテンポやコードも同時に推定する統一的なネットワークを設計することが挙げられる．注意重みのピーク位置は楽曲のテンポに依存し，小節線位置でコードが変化しやすい傾向があることから，複数の音楽要素の同時推定は互いの推定精度を向上させると期待される．

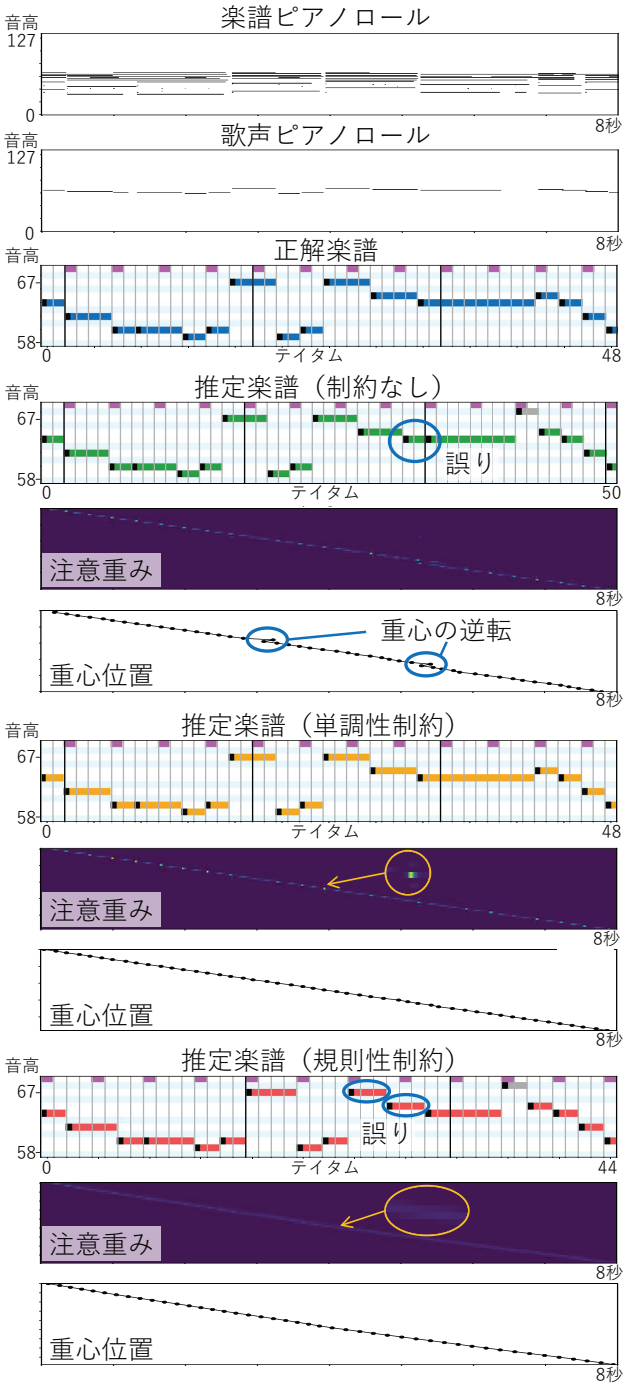


図 3　提案法を用いて推定された楽譜と注意重みの例．楽譜中の黒い四角は音符のオンセット，灰色の四角は休符，紫色の四角はビート，黒い縦線は小節線，灰色の縦線はテイタムを表す．

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

## 参考文献

[1] Luong, M.-T., Pham, H. and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, *EMNLP*, pp. 1412–1421 (2015).
[2] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *ICLR*, pp. 1–15 (2015).
[3] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-Based Models for Speech Recognition, *NIPS*, pp. 577–585 (2015).
[4] Chan, W., Jaitly, N., Le, Q. and Vinyals, O.: Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition, *ICASSP*, pp. 4960–4964 (2016).
[5] Prabhavalkar, R., Sainath, T. N., Li, B., Rao, K. and Jaitly, N.: An Analysis of "Attention" in Sequence-to-Sequence Models, *INTERSPEECH*, pp. 3702–3706 (2017).
[6] Nishikimi, R., Nakamura, E., Fukayama, S., Goto, M. and Yoshii, K.: Automatic Singing Transcription Based on Encoder-Decoder Recurrent Neural Networks with a Weakly-Supervised Attention Mechanism, *ICASSP*, pp. 161–165 (2019).
[7] Chiu, C. and Raffel, C.: Monotonic Chunkwise Attention, *ICLR* (2018).
[8] Molina, E., Tardón, L. J., Barbancho, A. M. and Barbancho, I.: SiPTH: Singing Transcription Based on Hysteresis Defined on the Pitch-Time Curve, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 23, No. 2, pp. 252–263 (2015).
[9] Kroher, N. and Gómez, E.: Automatic Transcription of Flamenco Singing from Polyphonic Music Recordings, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 24, No. 5, pp. 901–913 (2016).
[10] Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J. and Dixon, S.: Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency, *TENOR*, pp. 23–30 (2015).
[11] Yang, L., Maezawa, A., Smith, J. B. L. and Chew, E.: Probabilistic Transcription of Sung Melody Using a Pitch Dynamic Model, *ICASSP*, pp. 301–305 (2017).
[12] Benetos, E. and Dixon, S.: Multiple-Instrument Polyphonic Music Transcription Using a Temporally Constrained Shift-Invariant Model, *The Journal of the Acoustical Society of America*, Vol. 133, No. 3, pp. 1727–1741 (2013).
[13] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S. and Eck, D.: Onsets and Frames: Dual-Objective Piano Transcription, *ISMIR*, pp. 50–57 (2018).
[14] Wu, Y.-T., Chen, B. and Su, L.: Polyphonic Music Transcription with Semantic Segmentation, *ICASSP*, pp. 166–170 (2019).
[15] Raphael, C.: A Hybrid Graphical Model for Rhythmic Parsing, *Artificial Intelligence*, Vol. 137, No. 1-2, pp. 217–238 (2002).
[16] Hamanaka, M., Goto, M., Asoh, H. and Otsu, N.: A Learning-Based Quantization: Unsupervised Estimation of the Model Parameters, *ICMI*, pp. 369–372 (2003).
[17] Nakamura, E., Yoshii, K. and Sagayama, S.: 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).
[18] Nishikimi, R., Nakamura, E., Goto, M., Itoyama, K. and Yoshii, K.: Scale- and Rhythm-Aware Musical Note Estimation for Vocal F0 Trajectories Based on a Semi-Tatum-Synchronous Hierarchical Hidden Semi-Markov Model, *ISMIR*, pp. 376–382 (2017).
[19] Nakamura, E., Nishikimi, R., Dixon, S. and Yoshii, K.: Probabilistic Sequential Patterns for Singing Transcription, *APSIPA ASC*, pp. 1905–1912 (2018).
[20] Carvalho, R. G. C. and Smaragdis, P.: Towards End-to-End Polyphonic Music Transcription: Transforming Music Audio Directly to a Score, *WASPAA*, pp. 151–155 (2017).
[21] Román, M. A., Pertusa, A. and Calvo-Zaragoza, J.: An End-To-End Framework for Audio-To-Score Music Transcription on Monophonic Excerpts, *ISMIR*, pp. 34–41 (2018).
[22] Nienhuys, H.-W. and Nieuwenhuizen, J.: LilyPond, a System for Automated Music Engraving, *CIM*, pp. 167–171 (2003).
[23] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *NIPS*, pp. 3104–3112 (2014).
[24] Graves, A., Fernandez, S., Gomez, F. and Schmidhuber, J.: Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks, *ICML*, pp. 369–376 (2006).
[25] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases, *ISMIR*, pp. 287–288 (2002).
[26] Goto, M.: AIST Annotation for the RWC Music Database., *ISMIR*, pp. 359–360 (2006).
[27] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *arXiv preprint arXiv:1412.6980*, pp. 1–15 (2014).
[28] He, K., Zhang, X., Ren, S. and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *ICCV*, pp. 1026–1034 (2015).