Vol.2012-MUS-96 No.10
2012/8/10

楽譜の文脈自由２次元木構造表現に基づく多重音スペクトログラム生成モデルによる音響信号からの自動採譜

土屋 政人[1]　落合 和樹[1]　亀岡 弘和[1,2]　嵯峨山 茂樹[1]

概要：
本報告では音響信号からの自動採譜を目的とし，2 次元木構造によって楽譜が生成される高次プロセスと楽譜からスペクトログラムが生成される低次プロセスを一体化した生成モデルを構築し，モデルのパラメータを推論することで自動採譜を実現するアプローチを提案する．自動採譜は多重音中の各ノートの音高や発音時刻に加えてテンポや音価の推定といった様々な問題が相互依存する大変複雑な課題である．人間が音楽を理解する過程では，リズムやハーモニーといった音楽上の常識的な文法に当てはまるように多重音中の個々のノートを聴き分け，テンポやリズムを認識していると考えられる．このように，観測音響信号がどのようなノートの音で構成されているらしいか，各ノートがどういうテンポのもとでどのような楽譜が意図されて演奏されているらしいか，その楽譜がどの程度音楽的に常識的かどうか，という推論を協調的に行うことで自動採譜の問題が効果的に解決できる可能性がある．そこで本研究では，楽譜が 2 次元木構造によって生成されるプロセスと楽譜からスペクトログラムが生成されるプロセスを一体化した生成モデルを定式化し，観測音響信号からモデルのパラメータを推論することで採譜を行うアプローチを提案する．実演奏の音楽音響信号を対象とした特定条件下での音高推定・リズム解析の予備実験を行い，本手法の有効性を示した．

キーワード：自動採譜，確率文脈自由文法，多重音解析，変分ベイズ，確率的生成モデルアプローチ

Bayesian polyphonic spectrogram modeling based on context-free 2D tree structure representation of musical score for automatic music transcription

Tsuchiya Masato[1]　Ochiai Kazuki[1]　Kameoka Hirokazu[1,2]　Sagayama Shigeki[1]

Abstract: In order to transcribe scores from the polyphonic acoustic signals, we propose a generative model consisting following two processes; (1) generating score by using 2-dimentional tree structure, and (2) generating spectrograms from the score. We also establish an estimating algorithm for music transcription from the observed spectrograms. Music transcription task contains various interdependent subproblems such as multipitch analysis, onset detection, tempo estimation and rhythm estimation. The interdependency between these subproblems makes music transcription more difficult, and how ambiguity between tempo and rhythm should be dealt with is unsolved problem. There are grammatical rules about rhythm and harmony in Western tonal music. When we listen to the music, it seems that we unconsciously recognize the rhythm and tempo with the grammatical rules and resolve above-stated interdependency. In this way, there is a possibility of improving the result of transcription by solving these subproblems cooperatively. To solve this interdependent subproblems, we propose the integrated generative model based on the musical grammars. This model is designed to generate the scores and the spectrograms in a single step. We also propose the decoding method to infer the parameters from the observed spectrograms. The results of our preliminary experiment indicated that our system can transcribe music with fixed number of notes and fixed length of the input music.

Keywords: Music Transcription, Probabilistic Context Free Grammar, Multipitch Analysis, Variational Bayes, Probabilistic Generative Model Approach

1. はじめに

自動採譜とは人間が演奏した結果としての音楽信号を楽譜へと変換することを指す．自動採譜システムは即興音楽や絶版になった楽譜を作成したり，音楽データベースを楽譜化し記号ベースの音楽検索等に利用することができ，音楽音響信号処理における重要な課題のひとつである．従来の研究では単旋律の音響信号の楽譜化は達成されているが，多重音の楽譜化はいまだに十分な精度で実現されているとは言えず，研究の余地が多く残っている．

自動採譜には大きくわけて 2 つの解くべき課題がある．一つは音響信号から音高，オンセット時刻を推定するという多重音解析の問題，もう 1 つは多重音解析で推定されたオンセット時刻列からテンポや音価の推定を行うというリズム解析の問題である．これらが相互に関連しあうことで強い曖昧性を生み，自動採譜は達成困難な課題とされてきた．多重音解析の問題は楽器の単音スペクトログラムが調波性構造を持つことに起因する．多重音の場合は同時に演奏されるそれぞれの単音スペクトログラムが重ね合わさった状態で観測されるため，ピッチを正確に推定するのがさらに困難になる．一方でリズム解析の問題は人間の演奏が常にテンポの変動やオンセット時刻のゆらぎを含んでいることに因る．楽譜上の音価長の単位を Tick と定義すると，式 1 のようにある音符の実時間上での長さは楽譜上の拍長と演奏テンポを乗算した値である．

$$\text{実時間 (sec)} = \text{音価 (Tick)} \times \text{テンポ (sec/Tick)} \quad (1)$$

観測されるのは実時間上の値であるため，右辺の値には無限の組み合わせが存在する．よって，より自然な楽譜へ変換するにはよくつかわれるリズムパターンで，なおかつ極端なテンポ変動をしないような解釈をしなければならない．

これら 2 つの課題に対して，従来の研究では多段処理的なアプローチをしてきた [1], [7]．つまり，音響信号を多重音解析によってピッチ，オンセット時刻情報にした後にリズム解析によってテンポや音価推定を行っていた．しかし，人間が音楽を認識する際はリズムや和声に関する音楽として常識的な文法構造を意識し，それに当てはまるように個々のノートの音を聞き分けながらリズム・テンポを認識しているはずで，上に述べたような多段決定的なプロセスは人間が音楽を認識するシステムとは異なっていると考えられる．こうした互いに依存関係にある要素は鶏と卵の関係にあり，ある 1 つの要素が正しく推定されれば，その結果は他の要素の推定結果をより改善するため，これらは同時推定を行うことによってより精度よく推定できる可能性がある [6]．

そこで本研究では多段処理的に別々な処理するのではなく協調的に同時推論を行うために，楽譜とオンセットが 2 次元木構造によって生成される高次プロセスとその情報を元にスペクトログラムが生成される低次プロセスを一体化した生成モデルを提案する．そして，逆問題として音高，オンセット，楽譜のリズム，テンポといったモデルのパラメータを推論するアルゴリズムを構築し，実際の音楽音響信号からの楽譜推定精度を検証する実験を行う．

2. 階層的ベイズモデルによる音楽音響スペクトログラムの生成

2.1 モデルの概要

提案する生成モデルは大きく分けて以下に示す 3 つのサブプロセスからなる: (1) 二次元的な木構造に基づき楽譜を生成するプロセス，(2) テンポカーブを生成するプロセス，(3) 上記の (1),(2) のプロセスから生成された情報を元に実際の音響スペクトログラムを生成するプロセス．(1) と (2) は互いに独立であり，(3) は (1) と (2) に依存している．以下の小節でそれぞれ (1) は 2.2，(2) は 2.3，(3) は 2.4 で述べる．採譜を行う際は楽譜上の各音符の配置やオンセット時刻などをすべて確率変数としてパラメータ化し，入力音響信号のスペクトログラムを生成するようなパラメータをベイズ法を用いて推定を行う．これらのパラメータを推定するアルゴリズムに関しては第 3 節で述べる．

2.2 楽譜の生成プロセス

第 1 節で述べていた音楽の常識的な文法とは何を指すのか．楽譜上のリズムにはよくつかわれるリズムのパターンが存在する．リズムを言語処理における単語と捉えると，音楽の常識的な文法とはこの頻出のリズムパターンを意味する．自然言語と違うところは楽譜は横方向（時間方向）だけでなく，縦方向（音高方向）にも広がりを持っていることである．単旋律の音楽が作曲される時を考えると，音楽はリズムより大きいかたまりとしてモチーフやフレーズといった時間方向への階層構造を持っている．これを多重音の作曲へ拡張すると，多重音の場合はさらにボイシングなど音高方向に対する階層構造を持っており，実在する楽譜はこういった 2 次元の階層構造が複雑に組み合わさったものであるといえる．故に，楽譜は 2 次元木構造によって捉えることでこういった複雑な構造もトップダウン的に生成することができると考えられる．

図 1 に 2 次元木構造によって楽譜を生成するプロセスの一例を示す．この例では初めは全音符が 1 つだけある状態からスタートし，最初は時間方向への分割 (Time-Spanning) を行うことによって 2 つの 2 分音符へと分割する．次に左側の音符が音高方向にコピー (Synchronization) を行うことによって，多重音を生成する．最後に右側の音符がもう

[1] 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
[2] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories Media Information Laboratory

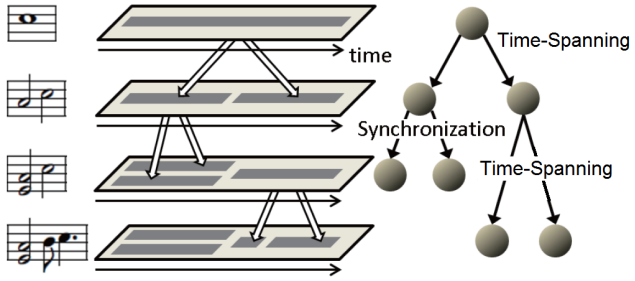

図 1 2 次元木構造表現による楽譜の生成

Fig. 1 Generative model of a 2-dimensional tree structure representation of musical notes.

全ての非終端記号の要素に適用する生成規則：

$\phi^T \sim \mathrm{Beta}(\phi^T; 1, \beta^T)$
（終端記号出力か分割か）

$\phi^N \sim \mathrm{Beta}(\phi^N; 1, \beta^N)$
（分割が時間方向か音高方向か）

$\phi^B_{l,l'} \sim \mathrm{Dirichlet}(\boldsymbol{\phi}^B_l; 1, \beta^B)$
（どの拍位置で分割するか）

導出木の全てのノードに適用する生成規則：

$b_n \sim \mathrm{Bernoulli}(b_n; \phi^T)$
（出力か分割かを選択）

If $b_n = (\mathrm{EMISSION})$
$S_r \sim \delta_{S_r, S_n}, \quad L_r \sim \delta_{L_r, L_n}$
（終端記号を出力）

If $b_n = (\mathrm{BINARY\text{-}PRODUCTION})$
$\rho_n \sim \mathrm{Bernoulli}(\rho_n; \phi^N)$
If $\rho_n = (\mathrm{SYNCHRONIZATION})$
$S_{n_1} \sim \delta_{S_{n_1}, S_n}, \quad S_{n_2} \sim \delta_{S_{n_2}, S_n}$
$L_{n_1} \sim \delta_{L_{n_1}, L_n}, \quad L_{n_2} \sim \delta_{L_{n_2}, L_n}$
（親ノードを複製）
If $\rho_n = (\mathrm{TIME\text{-}SPANNING})$
$S_{n_1} \sim \delta_{S_{n_1}, S_n}, \quad S_{n_2} \sim \delta_{S_{n_2}, S_n + L_{n_1}}$
$L_{n_1} \sim \delta_{L_{n_1}, L_n - L_{n_1}}$
$L_{n_2} \sim \mathrm{Categorical}(L_{n_2}; \boldsymbol{\phi}^B_l)$
（時間分割をして子ノードを生成）

図 2 2 次元木構造表現による楽譜生成モデルの確率的定式化．なお，δ は Kronecher's デルタであり，$x \sim \delta_{x,y}$ は $x = y$ ならば確率値 1 であり，それ以外は 0 を示す．

Fig. 2 The probabilistic formulation of generative model of a 2-dimensional tree structure representation. δ denotes Kronecker's delta. Thus, $x \sim \delta_{x,y}$ means $x = y$ (with probability 1).

一度 Time-Spanning を行うことによって 8 分音符と付点 4 分音符に分割され，最終的な楽譜が生成される．このプロセスは分割途中の音符を非終端記号とみた確率文脈自由文法 (PCFG) の 2 次元拡張版であるといえる．

このプロセスに対して PCFG をベイズモデルとして定式化した [5] に倣って確率的定式化を行うと図 2 のようになる．各ノードはまず第一の選択として分割するのをやめて終端記号を生成するか，分割して 2 つの子ノードを生成するかの選択を行い，そして，後者が選ばれた場合はさらに第二の選択として Synchronization を行うか，Time-Spanning を行うかの選択を行う必要がある．そこで，この 2 者選択を Bernoulli 分布に基づく確率変数によって決定することで確率的に捉えることができる．Time-Spanning が選択された場合は音符をどの長さに分割するか選択しなければならないので， Categorical 分布に基づく変数によってこれを決定する．Bernoulli 分布や Categorical 分布に渡すパラメータは計算の便宜上，それぞれの共役事前分布である，Beta 分布，Dirichlet 分布にした．[*1] この生成文法によって無数のリズムパターンを生成することができ，これらのパターンの使われやすさを確率的に扱うには各，確率変数の事前分布のパラメータによって制御すればよい．2 つの選択においてどちらが選択されやすいかは Beta 分布の超パラメータ ϕ^N, ϕ^T で調節することができ，さらに Time-Spanning の分割位置は Dirichlet 分布の超パラメータ ϕ^B によって特定の間隔で切れやすい傾向を持つように調節できる．

木構造における各ノードは開始拍位置 $S_{\boldsymbol{n}}$，長さ $L_{\boldsymbol{n}}$ という情報を持つ．初期状態は 1 曲全体の長さ D を持つ 1 つのノード $S_{\mathbf{1},\boldsymbol{R}} = 1, \quad L_{\mathbf{1},\boldsymbol{R}} = D$ が 1 つあり，これに対して図 2 に示す生成文法を再帰的に適用する．ここで，R は対象とする曲の音符数，$\boldsymbol{n}$ は CYK アルゴリズムにおける三角行列のインデックスであり，$\boldsymbol{n} = (1, R)$ はそのノードから 1〜R 番目の音符が生成されることを表す．

最終的に，図 2 中の 1 つ目の選択において $b_{\boldsymbol{n}}$ =EMISSION が選択された場合，拍時刻 S_n と拍長 L_n を，終端記号として出力する．音符の音高は調によって特定の音高が出やすい傾向があるため，音高 κ_r を以下のような分布に従って生成するようにすることで，特定の音高のでやすさをモデル化できる．

$$\kappa_r \sim \mathrm{Categorical}(\kappa_r; \phi^K_r) \tag{2}$$

$$\boldsymbol{\phi}^K \sim \prod_r \mathrm{Dirichlet}(\boldsymbol{\phi}^K_r; \boldsymbol{\alpha}^K_r) \tag{3}$$

ここで，$\mathrm{Categorical}(x; \boldsymbol{y}) = y_x, (\boldsymbol{y} = (y_1, \ldots, y_I), \sum_i y_i = 1)$ であり，$\mathrm{Dirichlet}(\boldsymbol{y}; \boldsymbol{z}) \propto \prod_i y_i^{z_i - 1} (\boldsymbol{z} = (z_1, \ldots, z_I))$ はその共役事前分布にあたる．

2.3 テンポの生成プロセス

人間が音楽を演奏する際，多くの場合は演奏速度が一定でなくゆるやかに変化している．曲に演奏表情を付けるために意図的に変化をつける場合でなくとも，人間の演奏は常に若干の速度変動を含み，機械のように完全に一定速度で演奏するのは不可能といえる．そこで楽譜上の先頭から数えた拍位置の d Tick での瞬間的なテンポ値を μ_d として，このテンポ値が直前の $d-1$ Tick での瞬間テンポ値 μ_{d-1} にのみ依存するという 1 次マルコフ連鎖を仮定すると，ゆ

*1 ベイズ推定法では計算をしやすくするために，共役事前分布を事前分布とするのが一般的である．

るやかなテンポ変動は正規分布を用いて以下のように定式化することができる．

$$p(\boldsymbol{\mu}) = \prod_{d=2}^{D} \mathcal{N}\left(\mu_d; \mu_{d-1}, (\sigma^{\mu})^2\right) \tag{4}$$

μ_d は楽譜上の $d-1$ Tick の位置から d Tick の位置までの実時間上の経過時間であるから，拍時刻 ψ_d を導入して以下のように書くこともできる．

$$p(\boldsymbol{\psi}|\boldsymbol{\mu}) = \prod_{d=2}^{D} \mathcal{N}\left(\psi_d; \psi_{d-1} + \mu_{d-1}, (\sigma^{\psi})^2\right) \tag{5}$$

2.4 音響スペクトログラムの生成プロセス

ここでは 2.2 と 2.3 で生成された楽譜上のオンセット位置とテンポ情報から実際の音響スペクトログラムを生成するプロセスについて述べる．このモデル化にあたっていくつかの仮定を置いた．

(1) 単音（単音モデル）のスペクトルは定常であり，ピッチにのみ依存する
(2) 音楽は様々な音高や音長をもつ単音の和で表現される
(3) 各音のパワーはオンセットからオフセットまで滑らかに連続的な形状を持つ

1 つ目と 2 つ目の仮定によって R 個の単音から構成される音響スペクトログラムは以下の形で表現することができる．

$$X_{\omega,t} = \sum_{r=1}^{R} H_{\omega,\kappa_r} W_{r,t} \tag{6}$$

ここで，ω, t はそれぞれ周波数，時間のインデックスである．H_{ω,κ_r} はピッチが κ_r である単音の定常なスペクトルを表しており，$W_{r,t}$ は r 番目のノートの時刻 t における信号のパワーを示している．仮定の 1 と 2 は必ずしも現実の観測スペクトログラム $Y_{\omega,t}$ に常に適用できるわけではなく，モデル化したスペクトログラム $X_{\omega,t}$ との間にはある程度のずれが存在する．そこで，このずれを表現するための 1 つの方法として $Y_{\omega,t}$ は $X_{\omega,t}$ を中心とする Poisson 分布に従うと仮定した．

$$Y_{\omega,t} \sim \text{Poisson}\left(Y_{\omega,t}; X_{\omega,t}\right) \tag{7}$$

なお，$\text{Poisson}(y; x) = x^y e^{-x}/y!$ として定義される Poisson 分布において $X_{\omega,t}$ に関して最大化を行うことは I ダイバージェンス基準を距離関数とした $X_{\omega,t}$ を $Y_{\omega,t}$ へフィッティングさせる最適化問題を解いていることに相当する．[4]

スペクトル基底 $H_{\omega,k}$ は調波性を持つような拘束が入るべきであるので，そのための 1 つの手段として $H_{\omega,k}$ がなんらかの事前分布に従うようにすることで実現することができる．今回は Gamma 分布によってモデル化を行った．

$$H_{\omega,k} \sim \text{Gamma}(H_{\omega,k}; \beta^H_{\omega,k}\bar{H}_{\omega,k} + 1, \beta^H_{\omega,k}) \tag{8}$$

ただし，$\text{Gamma}(x; a, b) \propto x^{a-1}e^{-bx}$ である．$\bar{H}_{\omega,k}$ は事前知識としてピッチ k が持つべきスペクトルの形状であり，β はスペクトルの最頻値周辺の尖度を調整する．

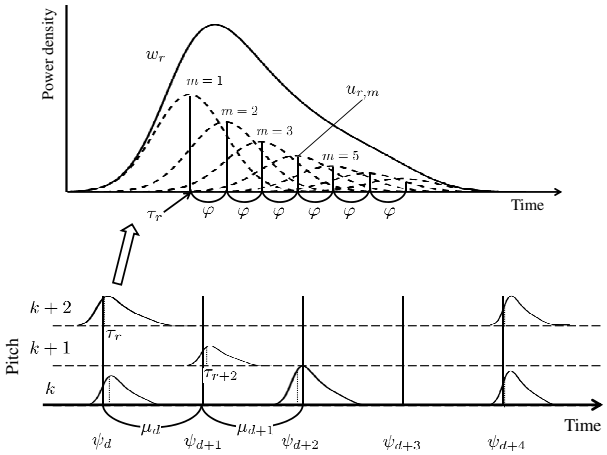

図 3 正規分布の混合数を $M = 7$ としたときの単音のパワーエンベロープモデル $W_{r,t}$（上）と単音モデルが配置された拍構造モデル（下）．

Fig. 3 Power envelop of a musical note (top) and rhythm structure with deployed power envelop of musical notes (bottom).

楽器音のパワーに関しては多くの場合，オンセットで急峻な上昇をした後でゆるやかに減少していくことが多いので，上に述べた 3 つ目の仮定を置いた上で，各音のパワーの時間変化 $W_{r,t}$ を，拘束つき混合正規分布型の関数 [4] によりモデル化する．

$$W_{r,t} = \sum_{m=1}^{M} G_{r,m,t} \tag{9}$$

$$= \sum_{m=1}^{M} \frac{w_r u_{r,m}}{\sqrt{2\pi}\varphi} e^{-\{t-(m-1)\varphi-\tau_r\}^2/2\varphi^2} \tag{10}$$

τ_r はオンセット時刻であり，図 3 に示すように τ_r はエンベロープを構成する先頭の正規分布の中心となっている．各混合正規分布の分散とその配置の間隔は φ という定数に統一した．w_r は r 番目のノートのパワーであり，必要最低限の単音モデルだけエネルギーが大きくなるようにするため，

$$w_r \sim \text{Gamma}(w_r; \alpha^w_r, \beta^w_r) \tag{11}$$

のように Gamma 分布に従うようにした上で，パラメータ α^w_r, β^w_r を調整することで $H_{\omega,k}$ と同様にスパース性を保つようにすることができる．

エンベロープを構成する混合正規分布の数 M は観測信号から推定されるべきである．$u_{r,1}, \ldots, u_{r,M}$ は各混合正規分布の重みであり，総和が 1 の単調減少数列としている．そこでこれらを生成するプロセスとして Dirichlet 過程の 1 つである Stick-Breaking Construction [8] を利用することができる．

$$u_{r,m} = V_{r,m} \prod_{m'=1}^{m-1} (1 - V_{r,m'}) \tag{12}$$

$$V_{r,m} \sim \text{Beta}(V_{r,m}; 1, \beta^V_{r,m}) \tag{13}$$

これによって m が大きくなるにつれて $u_{r,m}$ が小さくなっていくので，各ノートの減衰を表現することができる．

2.3 ではテンポ変動をモデル化したが，人間の演奏はそれに加え本来のオンセット時刻 ψ_{S_r} からのタイミングのずれがある．そこで，r 番目の単音のオンセット時刻 τ_r を，ψ_{S_r} を中心とした正規分布に従うようにした．

$$p(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{S}) = \prod_r \mathcal{N}\left(\tau_r;\psi_{S_r},(\sigma^\tau)^2\right) \tag{14}$$

ここで，ψ_{S_r} は 2.2 で述べた楽譜生成プロセスによって生成された r 番目の音符の拍開始位置 (Tick) を表す変数である．

2.5 潜在変数の導入

解析に変分ベイズ法を用いるため，その変分事後分布を計算する便宜上，潜在変数 $C_{r,m,\omega,t}$ を導入した．

$$C_{r,m,\omega,t} \sim \text{Poisson}\left(C_{r,m,\omega,t}; H_{\omega,\kappa_r} G_{r,m,\omega,t}\right) \tag{15}$$

$$Y_{\omega,t} \sim \delta\left(Y_{\omega,t} - \sum_{r,m} C_{r,m,\omega,t}\right) \tag{16}$$

この式は $C_{r,m,\omega,t}$ について周辺化することで式 7 に帰着することができる．

3. パラメータ推論

3.1 ベイズ推定による事後分布の推定

この章では提案した生成モデルのパラメータ推定の方法について論じる．ここで，観測信号 Y が与えられた時のパラメータ θ に関する事後分布 $p(\theta|\boldsymbol{Y})$ を推定したい．

$\theta = \{\boldsymbol{H},\boldsymbol{w},\boldsymbol{V},\boldsymbol{\tau},\boldsymbol{\kappa},\boldsymbol{\psi},\boldsymbol{\mu},\boldsymbol{S},\boldsymbol{L},\boldsymbol{\phi}^B,\phi^T,\phi^N,\boldsymbol{\phi}^K\}$ が今回のシステムで推定すべきパラメータであり，整理すると以下のようになる．

$\boldsymbol{H} = \{H_{\omega,k}\}_{\omega,k}$：ピッチ k のスペクトル基底
$\boldsymbol{w} = \{w_r\}_r$：$r$ 番目のノートのエネルギー
$\boldsymbol{V} = \{V_{r,m}\}_{r,m}$：$r$ 番目のノートのエンベロープを構成する係数
$\boldsymbol{\tau} = \{\tau_r\}_r$：$r$ 番目のノートのオンセット時刻
$\boldsymbol{\kappa} = \{\kappa_r\}_r$：$r$ 番目のノートのピッチ
$\boldsymbol{\psi} = \{\psi_d\}_d$：$d$ Tick の拍時刻
$\boldsymbol{\mu} = \{\mu_d\}_d$：$d$ Tick から $d+1$ Tick の拍間隔
$\boldsymbol{S} = \{S_r\}_r$：$r$ 番目のノートの拍位置
$\boldsymbol{L} = \{L_r\}_r$：$r$ 番目のノートの音価
$\boldsymbol{\phi^B},\boldsymbol{\phi^T},\boldsymbol{\phi^N},\boldsymbol{\phi^K}$：文法生成確率テーブル

事後分布 $p(\theta|Y)$ を求めるためには $p(Y)$ を計算する必要があるが，この計算は非常に複雑な積分計算を要し，解析的に求めることはできない．そこで，以下では変分ベイズ法に基づき，事後分布 $p(\theta|Y)$ の近似分布を得るための反復アルゴリズムを導出する．

3.2 変分ベイズ法による事後分布の近似計算

変分ベイズ法はベイズ推定における高速な近似解法の 1 つである．求めたい事後分布 $p(\theta|\boldsymbol{Y})$ を近似するための変分事後分布 $q(\theta)$ を新たに導入し，その KL ダイバージェンスを最小化するように反復計算を行う．

$$\underset{\theta}{\text{argmax}}\ \text{KL}(q(\theta)||p(\theta|\boldsymbol{Y})) \tag{17}$$

KL ダイバージェンスは 2 つの確率分布の差異を示しており，これを最小化することで $p(\theta|\boldsymbol{Y})$ を最もよく近似するような $q(\theta)$ を求めることができる．ここで，$q(\theta)$ を，$q()q()\ldots$ のように積の形で表せる関数クラスとすると，結局以下の式によってパラメータの分布を近似する計算に帰着する．

$$q(\theta_i) \propto p(\theta_i)\exp\left\langle \log p(\boldsymbol{Y},\boldsymbol{\theta})\right\rangle_{q(\boldsymbol{\theta}/\theta_i)} \tag{18}$$

ただし，$\langle ...\rangle_{q(\boldsymbol{\theta}/\theta_i)}$ は $\boldsymbol{\theta}$ の θ_i 以外のパラメータで期待値を取ることを表す．

結合分布 $p(\boldsymbol{Y},\theta)$ は 2.2, 2.3, 2.4 で述べてきたような各パラメータの依存性により，以下の形に分解することができる．

$$\begin{aligned} p(Y,\theta) =& p(\boldsymbol{H},\boldsymbol{w},\boldsymbol{V},\boldsymbol{\tau},\boldsymbol{\kappa},\boldsymbol{\psi},\boldsymbol{\mu},\boldsymbol{S},\boldsymbol{L},\boldsymbol{\phi}^B,\phi^T,\phi^N,\boldsymbol{\phi}^K|\boldsymbol{Y}) \\ \propto& p(\boldsymbol{Y}|\boldsymbol{C})p(\boldsymbol{C}|\boldsymbol{H},\boldsymbol{w},\boldsymbol{V},\boldsymbol{\tau},\boldsymbol{\kappa})p(\boldsymbol{H})p(\boldsymbol{V})p(\boldsymbol{w}) \\ &\cdot p(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{S})p(\boldsymbol{\psi}|\boldsymbol{\mu})p(\boldsymbol{\mu})p(\boldsymbol{\kappa}|\boldsymbol{\phi^K})p(\boldsymbol{\phi^K}) \\ &\cdot p(\boldsymbol{S},\boldsymbol{L}|\boldsymbol{\phi}^B,\phi^T,\phi^N)p(\boldsymbol{\phi}^B)p(\phi^T)p(\phi^N) \end{aligned} \tag{19}$$

各パラメータの変分事後分布の更新式は以下のとおりとなる．

$$q(\boldsymbol{C}_{\omega,t}) = \text{Multinomial}(\boldsymbol{C}_{\omega,t};Y_{\omega,t},\boldsymbol{f}^C_{\omega,t}) \tag{20}$$
$$q(H_{\omega,k}) = \text{Gamma}(H_{\omega,k};\xi^H_{\omega,k},\zeta^H_{\omega,k}) \tag{21}$$
$$q(w_r) = \text{Gamma}(w_r;\xi^w_r,\zeta^w_r) \tag{22}$$
$$q(V_{r,m}) = \text{Beta}(V_{r,m};\xi^V_{r,m},\zeta^V_{r,m}) \tag{23}$$
$$q(\boldsymbol{\tau},\boldsymbol{\psi},\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\chi};\boldsymbol{\xi}^\chi,\boldsymbol{\zeta}^\chi) \tag{24}$$
$$q(\kappa_r) = \text{Categorical}(\kappa_r;\boldsymbol{f}^\kappa_r) \tag{25}$$
$$q(\boldsymbol{\phi}^K_r) = \text{Dirichlet}(\boldsymbol{\phi}^K_r;\boldsymbol{\xi}^K_r) \tag{26}$$
$$q(S_r,L_r) = \text{Categorical}(S_r,L_r;\boldsymbol{f}^{SL}_r) \tag{27}$$
$$q(\boldsymbol{\phi}^B_l) = \text{Dirichlet}(\boldsymbol{\phi}^B_l;\boldsymbol{\xi}^B_l) \tag{28}$$
$$q(\phi^T) = \text{Beta}(\phi^T;\xi^T,\zeta^T) \tag{29}$$
$$q(\phi^N) = \text{Beta}(\phi^N;\xi^N,\zeta^N) \tag{30}$$

ここで，$\boldsymbol{\chi},\boldsymbol{\xi}^\chi,\boldsymbol{\zeta}^\chi$ はそれぞれ添え字に書かれていない各要素を並べたベクトルで，

$$\boldsymbol{\chi} = \begin{bmatrix}\boldsymbol{\tau}\\ \boldsymbol{\psi}\\ \boldsymbol{\mu}\end{bmatrix},\ \boldsymbol{\xi}^\chi = \begin{bmatrix}\boldsymbol{\eta}^\tau\\ \boldsymbol{\eta}^\psi\\ \boldsymbol{\eta}^\mu\end{bmatrix},\ \boldsymbol{\zeta}^\chi = \begin{bmatrix}\boldsymbol{\nu}^\tau & \boldsymbol{\nu}^{\tau\psi} & \boldsymbol{\nu}^{\tau\mu}\\ \boldsymbol{\nu}^{\tau\psi} & \boldsymbol{\nu}^\psi & \boldsymbol{\nu}^{\psi\mu}\\ \boldsymbol{\nu}^{\tau\mu} & \boldsymbol{\nu}^{\psi\mu} & \boldsymbol{\nu}^\mu\end{bmatrix} \tag{31}$$

である．更新式を導出する際は，式 18 によって対数同時分布 $\log p(\boldsymbol{Y}, \theta)$ と対数変分事後分布 $\log q(\theta)$ を比較し，十分統計量を直接推定する．ここで，式 (27) から (30) は Inside-Outside アルゴリズムによって導くことができる．今回は紙面の都合上，導出された更新式は省略する．

4. 実験

提案手法の採譜精度を検証するために，音符数が既知の状態で実際の音楽音響信号から楽譜を推定できるかどうか実験を行った．これは音符数未知の場合に反復推定の過程で w_r が微量のエネルギーしかもたないものを排除し，正しい音符数にまで減少した状態から推定が行えるかという実験である．

実験データとして RWC クラシック音楽データベース・ジャズ音楽データベース [3] からピアノ曲（RWC-MDB-C-2001 No. 26, 27, 30）を選び，計算量の都合上，各曲最初の数小節分（約 10 s 程度）をモノラル信号にミックスダウンして 16 kHz にしたものを用いた．解析には Wavelet 変換されたスペクトログラムを用い，その条件は，時間分解能 16 ms，最低周波数 30 Hz，周波数分解能 12 cent とした．モデルの各事前分布のパラメータや，求めるパラメータの初期値は，$K = 74, M = 40, \varphi = 3, \alpha^H_{\omega,k} = \beta^H_{\omega,k}\bar{H}_{\omega,k} + 1, \beta^H_{\omega,k} = 500, \alpha^w_r = \beta^w_r = 0, \beta^V_{r,m} = 10e^{-m/8}/\sum_{m'} e^{-m'/8}, \sigma^\tau = 2, \sigma^\psi = 1, \sigma^\mu = 0.5, \alpha_{r,k} = 2, \beta^T = 1, \beta^N = 2$ とした．$\bar{\boldsymbol{H}}$ は基底スペクトル $\boldsymbol{H}$ の初期値であり，RWC 楽器音データベース [2] のピアノの単音データに対して，非負値行列分解 [9] を適用して求めたものとした．パラメータ更新の反復回数は 10 回とした．変分ベイズによって楽譜やオンセットに関するパラメータの事後分布が推定された後はそれらの期待値を計算することで実際に推定されたパラメータとした．

採譜結果の一部を図 4 に示す．このようなテンポ変動が少ない曲の場合はリズムよりも音高に関して多くのエラーが見られた．これは，低次のスペクトログラム生成プロセスに関するエラーであり，2.4 で置いた仮定が正しくない可能性がある．又，テンポ変動が大きく，早い曲に関してはリズム誤りが増加する傾向が見られ，テンポ変動のモデル化にもまだ改良が必要であるといえる．

5. おわりに

本報告では音高推定，オンセット時刻推定，テンポ推定，リズム推定という多くの課題を含む自動採譜の課題に対して，楽譜生成から音響スペクトログラム生成まで統一された 1 つの生成モデルによって捉えることでこれらを同時推定する手法を提案した．実験では音符数が既知の状態での音楽音響信号からの採譜結果を示した．今後の課題としてはより音楽的な知識を取り入れた生成文法によって楽譜をモデル化することや，現在は手動で与えている文法適用確率を実在する楽譜情報から学習すること，そして，オンセットに加えてオフセットも生成することで多重音の音価も正しく推定できるようにすることなどが挙げられる．

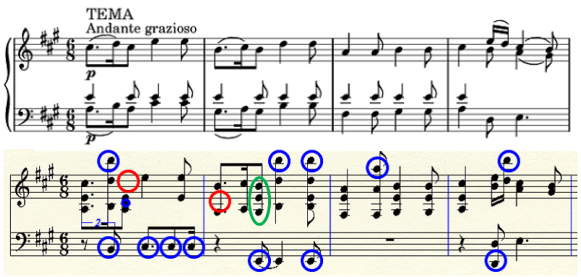

図 4 正解楽譜（上）提案手法で推定された楽譜（下）．赤，緑，青の円はそれぞれ消失誤り，音高誤り，オクターブ誤りを示す．

Fig. 4 Original score (top) and estimated score (bottom). the red, green and blue circles indicate the deletion errors, pitch errors and octave errors, respectively.

謝辞 本研究の一部は，文部科学省/学術振興会科学研究補助費 課題番号 (23240021) から補助を受けて行われた．

参考文献

[1] N. Bertin, R. Badeau, and G. Richard. Blind signal decompositions for automatic transcription of polyphonic music: Nmf and k-svd on the benchmark. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol. 1, pp. I–65. IEEE, 2007.

[2] M. Goto, et al. Development of the rwc music database. In *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, Vol. 1, pp. 553–556, 2004.

[3] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical, and jazz music database. In *Proc. ISMIR*, pp. 287–288, October 2002.

[4] Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 982–994, March 2007.

[5] P. Liang, S. Petrov, M. Jordan, and D. Klein. The infinite pcfg using hierarchical dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 688–697, 2007.

[6] K. Ochiai, H. Kameoka, and S. Sagayama. Explicit beat structure modeling for non-negative matrix faxtorization-based multipitch analysis. Proc. ICASSP’ 12, 2012.

[7] Stanislaw Andrzej Raczyński, Nobutaka Ono, and Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proc. ISMIR*, pp. 381–386, September 2007.

[8] J. Sethuraman. A constructive definition of dirichlet priors. Technical report, DTIC Document, 1991.

[9] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. WASPAA*, pp. 177–180, October 2003.