社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[ポスター講演] 歌声 F0 軌跡に対する自動採譜のための準ビート同期セグメンタル HMM

錦見　亮†　中村　栄太†　糸山　克寿†　吉井　和佳†

† 京都大学 大学院情報学研究科
E-mail: †{nishikimi,enakamura,itoyama,yoshii}@sap.ist.i.kyoto-u.ac.jp

あらまし　本稿では連続的な歌声 F0 軌跡から離散的な音符系列を推定する統計的手法を示す．従来，音楽音響信号からフレームレベルの歌声 F0 を推定するための研究が多く行われているので，我々は記号的な楽譜の獲得を目的とした F0 軌跡からの音符推定に取り組む．音符推定に対する素朴なアプローチは，一定時間単位（例えば，半ビート）ごとに歌声 F0 を半音レベルで離散化する手法である．しかしながら，このアプローチでは歌声 F0 軌跡が楽譜に記された音高から大きく逸脱している場合にうまくいかない．音符のオンセットはビート時刻から遅れたり進んだりする（オンセット変動）うえ，歌声 F0 は歌唱表現により変動する（周波数変動）．これらの逸脱を扱うために，我々は音符がビート時刻にゆるく同期して変化するベイジアン隠れマルコフモデルを提案する．音符の半音レベルの音高とオンセット変動は潜在変数とみなし，周波数変動は出力分布によって記述する．音符，オンセット変動，周波数変動はギブスサンプリングを用いて同時に推定する．実験結果よりベースライン手法に対して提案手法の音符推定精度が向上したことが示せた．

キーワード　歌声, 自動採譜, セグメンタル隠れマルコフモデル

1. はじめに

歌声は通常ポピュラー音楽のメロディラインを形成しており，楽曲の雰囲気や印象に大きな影響を与えるため，歌声解析は音楽情報処理の分野において重要なテーマの一つである．歌声解析において，音楽音響信号に対する基本周波数（F0）推定 [1–7] や歌声分離 [8, 9] は広く研究されている．これらの技術は歌手同定 [10, 11] や歌声抑圧に基づくカラオケシステム [12, 13]，ユーザーが特定の音楽的要素に焦点を当てながら，より深く音楽を理解することを助ける音楽鑑賞システム [14] などに用いられている．

本研究では，歌声 F0 軌跡から音符系列を取り出すことを目的とした音符推定問題に取り組む．歌声の F0 推定に関する研究は多く行われているが，実用面では楽譜のような離散的（記号的）な情報を抽出する問題に取り組む必要がある．もし，ビート情報が既知であれば，この問題に対する素朴なアプローチは一定時間単位（例えば，半ビート）ごとに多数決によって半音レベルの F0 に離散化する方法である [14]．しかしながら，歌声の F0 が楽譜に記された正確な半音レベルの F0 から大きく逸脱する場合や，音符のオンセットが正確なビート時刻から大きく進んだり遅れたりするような歌い方での場合に，この方法ではしばしばうまくいかない．

この問題を解決するために，隠れた音符系列からどのように歌声 F0 軌跡が生成されるかを表現する隠れマルコフモデル（HMM）に基づく統計的手法を提案する (図 1)．楽譜に記された音符の F0 は半音間隔の離散的な値のみを取り，ビート，半ビート，1/4 ビートの位置で変わりやすい．一方で，実際の

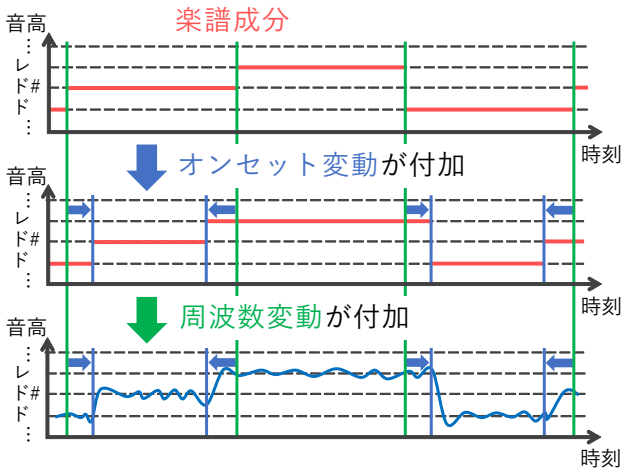


図 1: 歌声 F0 軌跡の生成過程．

歌声 F0 軌跡は連続的な信号であり，時間とともに動的に変化する．生成的な視点からこれら 2 種類の F0 を扱うために，メロディを歌った連続的な F0 が楽譜に書かれた離散的な F0 から時間方向と周波数方向に逸脱することを許した，準ビート同期セグメンタル HMM（Semi-beat-synchronous Segmantal HMM，SBS-SHMM）を提案する．提案する HMM において，音符の半音レベルの F0 とオンセット変動は潜在変数として表現され，音符の周波数変動は出力確率分布によってモデル化される．F0 軌跡とビート時刻が与えられれば，全ての変数と分布はギブスサンプリングを用いて同時に推定される．

2. 関 連 研 究

本章では歌声に関する関連研究を紹介する．

2.1 歌声音高推定

音楽音響信号に対する歌声 F0 軌跡の推定に関しては多くの研究がなされている [1–7]．Subharmonic Summation（SHS）[1] は基本周波数の候補 $\{f_0, \ldots, f_M\}$ のそれぞれについて高調波成分のパワーの総和を計算することで，各時刻ごとの基本周波数を決定する方法である．PreFEst [3] は多声音楽音響信号から最も優性な調波構造を抽出することによりメロディとベースラインの F0 軌跡を推定する手法である．池宮ら [2] は歌声分離と F0 推定を相補的に行う手法を提案している．まず，音楽音響信号に対する短時間フーリエ変換（STFT）によって得られたスペクトログラムからロバスト主成分分析（RPCA）[9] を用いて歌声が分離される．そして，分離した歌声に対して計算される SHS を用いてビタビ探索により歌声 F0 軌跡を求める．Salamon ら [4] はメロディ抽出に対して歌声 F0 軌跡の特徴を用いている．Durrieu ら [5] は，主旋律はソースフィルターモデルで表現され，伴奏音は非負値行列因子分解（NMF）に基づくモデルで表現されるメロディ抽出法を提案している．de Cheveigné ら [6] は基本周波数推定に対して誤り率を減少させるよう拡張した自己相関に基づく YIN と呼ばれる手法を提案している．Mauch ら [7] は確率的手法を用いて YIN を拡張することで複数の音高候補を出力するようにした pYIN と呼ばれる手法を提案している．

2.2 歌声音符推定

歌声 F0 軌跡の音高を離散化して音符系列を推定する手法も提案されている．1. 章で述べた多数決法は Songle [14] に実装されている．この手法では，歌唱表現や音高の生成過程について考慮していないので限界がある．Paiva ら [15] は 5 つの段階を経て多声音楽信号からメロディの音符を推定する手法を提案している．Raphael [16] は HMM に基づいて独唱歌声音響信号からリズム，テンポ，音符を同時に推定する手法を提案している．Poliner ら [17] は，音高はある基本周波数の高調波の集合として実現されるという仮定が必要がないサポートベクターマシン（SVM）に基づく手法を提案している．Laaksonen は [18] コード情報を用いてメロディ採譜手法を提案している．Ryynänen ら [19] はメロディ，ベースライン，コードを多声音楽から推定する手法を提案している．Mauch ら [20] によって開発された Tony というソフトウェアツールは HMM のビタビ探索によって pYIN が出力した複数の音高候補の中から音符を推定する．

2.3 歌声 F0 軌跡の解析

歌声 F0 軌跡から歌い方の個性や癖を抽出する研究もおこなわれている．大石ら [21] は時間変動と周波数変動を考慮した歌声 F0 軌跡の生成過程を表現するモデルを提案した．このモデルにおいて，歌声 F0 軌跡はノート，表現，微細変動の 3 つの成分から成る．ノート成分はノートの立ち上がりやオーバーシュートを含み表現成分はビブラートやポルタメントを含む．ノート成分と表現成分はノート指令と表現指令によって駆動される二次系線形システムの出力として表される．ノート指令と表現指令はそれぞれ音符系列と音楽的表現意図を表し，HMM を用いてモデル化される．この手法により歌声 F0 軌跡から歌唱表現の個人性を取り出すことは可能だが，事前に楽譜が与えらることを仮定しており，直接的に音符推定問題には適用できない．

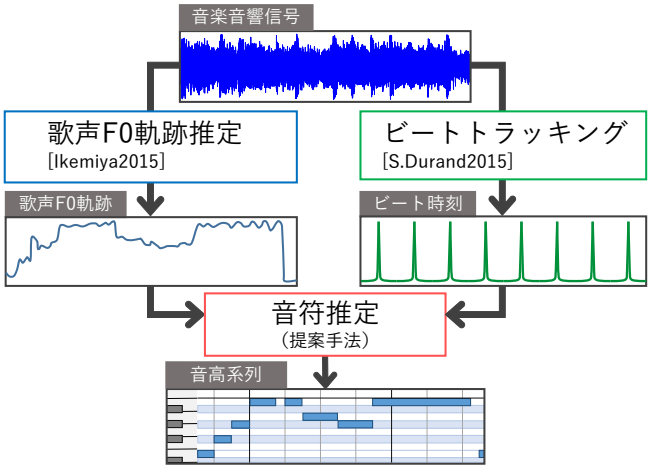


図 2: 提案する準ビート同期セグメンタル HMM に基づく音高推定手法の全体図．

3. 提 案 法

本章では準ビート同期セグメンタル HMM（SBS-SHMM）により歌声 F0 軌跡の生成過程を表現することで，F0 軌跡の背後にある音符系列を推定するための手法について説明する．観測として得られる歌声 F0 軌跡は，楽譜（音符系列）にオンセット変動と周波数変動が確率的に付与されて生成されるものとして定式化する．

3.1 問 題 設 定

音符推定問題を以下のように定める（図 2）．

入力: 歌声 F0 軌跡 $\boldsymbol{X} = \{x_t\}_{t=1}^{T}$ と音楽音響信号から自動推定された 16 分音符単位のビート時刻 $\boldsymbol{\psi} = \{\psi_n\}_{n=1}^{N}$
出力: 音高系列 $\boldsymbol{Z} = \{z_n\}_{n=1}^{N}$.

ここで t は時間フレームのインデックス，T は入力の F0 軌跡における時間フレームの数，x_t はフレーム t での対数周波数（単位はセント），N はビート時刻の数，ψ_n は n 番目のビート時刻（単位はフレーム），$z_n \in \{\mu_1, \ldots, \mu_K\}$ は ψ_{n-1} と ψ_n との間の音高であり，K は楽譜に現れる音高の種類の数である．曲の最初と最後はそれぞれ $\psi_0 = 1$ と $\psi_N = T{+}1$ で表される．本稿では，簡単化のため $z_n (n = 1, \ldots, N)$ はそれぞれ 16 分音符に対応するとする．16 分音符よりも長い音符は連続した同じ音高をもつ $\{z_n\}_{n=1}^{N}$ の系列で表される．

3.2 モデルの定式化

音符の音高遷移，オンセット変動，周波数変動を同時に表現する SBS-SHMM を定式化する．

3.2.1 音高遷移のモデル化

潜在音高系列 $\boldsymbol{Z}$ は以下に示すように一次マルコフ連鎖を

なす．

$$z_n|z_{n-1}, \boldsymbol{A} \sim \mathrm{Categorical}(z_n|\boldsymbol{a}_{z_{n-1}}) \tag{1}$$

ここで，$\boldsymbol{A} = [\boldsymbol{a}_1^T, \cdots, \boldsymbol{a}_K^T]$ は $K \times K$ の遷移確率行列であり，任意の $j \in \{1, \ldots, K\}$ について $\sum_{k=1}^{K} a_{jk} = 1$ が成り立つ．最初の潜在状態 z_1 は

$$z_1|\boldsymbol{\pi} \sim \mathrm{Categorical}(z_1|\boldsymbol{\pi}) \tag{2}$$

により決定される．ここで，$\boldsymbol{\pi} = [\pi_1, \cdots, \pi_K]^T$ は K 次元ベクトルであり，$\sum_{k=1}^{K} \pi_k = 1$ が成り立つ．

3.2.2 オンセット変動のモデル化

音符のオンセット変動 $\boldsymbol{\tau} = \{\tau_n\}_{n=1}^{N}$ は $\{-G, -G+1, \ldots, G-1, G\}$ の整数値を取る離散潜在変数として表される．n 番目の音符に対応する歌声のオンセット時刻を $\phi_n = \psi_n + \tau_n$ とおく．歌声 F0 軌跡の最初と最後では $\tau_0 = 0$，$\tau_N = 0$ とする．τ_n は以下のように確率的に生成される．

$$\tau_n|\boldsymbol{\rho} \sim \mathrm{Categorical}(\tau_n|\boldsymbol{\rho}) \tag{3}$$

ここで，$\boldsymbol{\rho} = [\rho_{-G}, \ldots, \rho_G]^T$ は $(2G+1)$ 次元ベクトルであり，$\sum_{g=-G}^{G} \rho_g = 1$ が成り立つ．

3.2.3 周波数変動のモデル化

観測 F0 x_t $(\phi_{n-1} \leqq t < \phi_n)$ は各ビート区間に割り当てられた半音単位の音高に確率的な周波数変動を付与することで生成される．x_t は各フレームにおいて独立に生成されるとし，n 番目のビート区間の出力確率は，以下のように与えられる．

$$b_{z_n \tau_{n-1} \tau_n} \equiv \left\{ \prod_{t=\phi_{n-1}}^{\phi_n - 1} p(x_t|z_n) \right\}^{\frac{1}{\phi_n - \phi_{n-1}}} \tag{4}$$

ここで，$p(x_t|z_n)$ は各フレームにおける出力確率である．遷移確率と出力確率のバランスを取るため，各フレームごとの出力確率の積をビート区間内のフレーム数で冪根をとる．我々は $p(x_t|z_n)$ にコーシー分布を用いる．コーシー分布は外れ値に頑健であり，以下のように

$$\mathrm{Cauchy}(x; \mu, \lambda) = \frac{\lambda}{\pi \left\{ (x-\mu)^2 + \lambda^2 \right\}} \tag{5}$$

で定義される．ここで，μ は位置パラメータであり，分布の中央値を決定する．また，λ は尺度パラメータである．n 番目のビート区間の音高が $z_n = k$ である時，μ は値 μ_k を取る．尺度パラメータは音高 z_n に依存しない値を取る．

実際の歌声 F0 軌跡は楽譜に記された音高から大きく逸脱することがあるので，コーシー分布の尺度パラメータは隣接する F0 の差 $\Delta x_t \equiv x_t - x_{t-1}$ に応じて変化するようにする．尺度パラメータは Δx_t の絶対値に比例するようにし，各フレームごとに以下のように定義する．

$$\lambda_t = c|\Delta x_t| + d \tag{6}$$

ここで，c は比例係数である．また，$\Delta_{x_t} = 0$ のとき $p(x_t|z_n)$ が計算できなくなる問題を避けるために変数 $d > 0$ を導入する．

3.3 事前分布の導入

我々はモデルパラメータ $\boldsymbol{A}$，$\boldsymbol{\pi}$，$\boldsymbol{\rho}$ に対して以下のようにディリクレ共役事前分布をおく．

$$\boldsymbol{a}_j \sim \mathrm{Dirichlet}(\boldsymbol{a}_j|\boldsymbol{\xi}_j) \tag{7}$$

$$\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\zeta}) \tag{8}$$

$$\boldsymbol{\rho} \sim \mathrm{Dirichlet}(\boldsymbol{\rho}|\boldsymbol{\eta}) \tag{9}$$

ここで，$\boldsymbol{\xi}_j = [\xi_{j1}, \ldots, \xi_{jK}]^T$ と $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_K]^T$ は K 次元ベクトル，$\boldsymbol{\eta} = [\eta_{-G}, \ldots, \eta_G]^T$ は $(2G+1)$ 次元ベクトルである．

また，コーシー分布の非負パラメータ c と d に対して以下のようにガンマ事前分布をおく．

$$c \sim \mathrm{Gamma}(c|c_0, c_1) \tag{10}$$

$$d \sim \mathrm{Gamma}(d|d_0, d_1) \tag{11}$$

ここで，c_0，d_0 は形状パラメータ，c_1，d_1 はレートパラメータである．

3.4 ベイズ推定

ベイズ推定の目的は事後分布 $p(\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{A}, \boldsymbol{\pi}, \boldsymbol{\rho}, c, d|\boldsymbol{X})$ を計算することである．この計算を解析的に行うのは難しいので，我々はマルコフ連鎖モンテカルロ（MCMC）法を用いてこれらの変数の値をサンプリングする．ここで，モデルパラメータを $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{\pi}, \boldsymbol{\rho}\}$ とする．$\boldsymbol{\Theta}$ のサンプルにはギブスサンプリングアルゴリズムを用いる．また，潜在変数系列 $\boldsymbol{Z}$，$\boldsymbol{\tau}$ のサンプルにはブロック化ギブスサンプリングアルゴリズムの一つである，フォワードフィルタリング・バックワードサンプリングアルゴリズムを用いる．これらのパラメータや変数は，HMM の教師なし学習で用いられるバウム・ヴェルチアルゴリズムとよばれる期待値最大化法（EM アルゴリズム）と同様の方法で反復更新される．c と d に関する分布は共役性が満たされないので，メトロポリス・ヘイスティングス（MH）アルゴリズムを用いて更新する．

3.4.1 潜在変数 Z，τ の推論

潜在変数系列 $\boldsymbol{Z}$ と $\boldsymbol{\tau}$ をサンプルする方法を説明する．各ビート区間ごとに，以下で与えられる確率を計算する．

$$\beta_{z_n \tau_n} = p(z_n, \tau_n | z_{n+1:N}, \tau_{n+1:N}, x_{1:T}) \tag{12}$$

ここで，$z_{n+1:N}$，$\tau_{n+1:N}$，$x_{1:T}$ はそれぞれ $z_{n+1}, \ldots, z_N$，$\tau_{n+1}, \ldots, \tau_N$，$x_1, \ldots, x_T$ を表す．n 番目のビート区間の潜在変数 (z_n, τ_n) は $\beta_{z_n \tau_n}$ に従いサンプルされる．式 (12) の計算と潜在変数のサンプリングはフォワードフィルタリング・バックワードサンプリングを用いて行われる．

フォワードフィルタリングでは以下の確率を反復計算で求める．

$$\alpha_{z_n \tau_n} = p(X_{0\tau_1}, \ldots, X_{\tau_{n-2}\tau_{n-1}}, X_{\tau_{n-1}\tau_n}, z_n, \tau_n)$$

ここで，$X_{\tau_{n-1}\tau_n}$ は $\phi_{n-1} = \psi_{n-1} + \tau_{n-1}$ から $\phi_n = \psi_n + \tau_n$ までのビート区間内の観測 x_t を表す．$\alpha_{z_n \tau_n}$ は以下のように計算される．

$$
\begin{aligned}
\alpha_{z_n\tau_n} &= p(X_{0\tau_1}, z_1, \tau_1) \\
&= p(X_{0\tau_1}|z_1, \tau_1)p(z_1)p(\tau_1) \\
&= b_{z_1 0\tau_1}\pi_{z_1}\rho_{\tau_1} \qquad (13) \\
\alpha_{z_n\tau_n} &= p(X_{0\tau_1}, \ldots, X_{\tau_{n-1}\tau_n}, z_n, \tau_n) \\
&= \sum_{\tau_{n-1}=-G}^{G} p(X_{\tau_{n-1}\tau_n}|z_n, \tau_{n-1}, \tau_n) \\
&\quad \cdot \sum_{z_{n-1}=1}^{K} p(X_{0\tau_1}, \ldots, X_{\tau_{n-2}\tau_{n-1}}, z_{n-1}, \tau_{n-1}) \\
&\quad \cdot p(z_n|z_{n-1})p(\tau_n) \\
&= \sum_{\tau_{n-1}=-G}^{G} b_{z_n\tau_{n-1}\tau_n} \sum_{z_{n-1}=1}^{K} \alpha_{z_{n-1}\tau_{n-1}} a_{z_{n-1}z_n}\rho_{\tau_n} \qquad (14)
\end{aligned}
$$

バックワードサンプリングでは，n 番目のビート区間において $\alpha_{z_n\tau_n}$ の値を用いて式 (12) が計算され，状態 (z_n, τ_n) が再帰的にサンプルされる．$(n{+}1)$ 番目の状態 (z_{n+1}, τ_{n+1}) がサンプルされた時，$\beta_{z_n\tau_n}$ は以下のように計算される．

$$
\begin{aligned}
\beta_{z_n\tau_n} &\propto p(X_{\tau_n\tau_{n+1}}|z_{n+1}, \tau_n, \tau_{n+1}) \\
&\quad \cdot p(z_{n+1}|z_n)p(\tau_{n+1}) \\
&\quad \cdot p(X_{0\tau_1}, \ldots, X_{\tau_{n-1}\tau_n}, z_n, \tau_n) \\
&= b_{z_{n+1}\tau_n\tau_{n+1}} a_{z_n z_{n+1}} \rho_{\tau_{n+1}} \alpha_{z_n\tau_n} \qquad (15)
\end{aligned}
$$

特に，潜在変数 (z_N, τ_N) は以下のように $\alpha_{z_N\tau_N}$ にしたがってサンプルされる．

$$
\beta_{z_N\tau_N} \propto \alpha_{z_N\tau_N} \qquad (16)
$$

3.4.2 モデルパラメータ A，π，ρ の学習

本章では $\boldsymbol{\Theta}$ の値の学習について説明する．バックワードサンプリングにおいてサンプルされた潜在変数の系列 $\{z_n, \tau_n\}_{n=1}^{N}$ について，$z_n = j$ かつ $z_{n+1} = k$ である遷移の数を s_{jk}，$\tau_n = g$ であるオンセット変動の数を u_g で表す．また，$z_1 = k$ の時，v_k の値を 1，それ以外を 0 とする．パラメータ a_{jk}，ρ_g，π_k は以下で与えられる事後分布からサンプルされることで更新される．

$$
p(\boldsymbol{a}_j|\boldsymbol{\xi}_j + \boldsymbol{s}_j) = \mathrm{Dirichlet}(\boldsymbol{a}_j|\boldsymbol{\xi}_j + \boldsymbol{s}_j) \qquad (17)
$$

$$
p(\boldsymbol{\rho}|\boldsymbol{\eta} + \boldsymbol{u}) = \mathrm{Dirichlet}(\boldsymbol{\rho}|\boldsymbol{\eta} + \boldsymbol{u}) \qquad (18)
$$

$$
p(\boldsymbol{\pi}|\boldsymbol{\zeta} + \boldsymbol{v}) = \mathrm{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\zeta} + \boldsymbol{v}) \qquad (19)
$$

ここで，$\boldsymbol{s}_j = [s_{j1}, \ldots, s_{jK}]^T$，$\boldsymbol{u} = [u_{-G}, \ldots, u_G]^T$，$\boldsymbol{v} = [v_1, \ldots, v_K]^T$ である．

3.4.3 コーシー分布のパラメータ c，d の学習

パラメータ c と d の推定には，MH アルゴリズムを用いる．コーシー分布は共役事前分布を持たないので c と d の事後分布を解析的に計算するのは困難である．c と d の値がそれぞれ c_i と d_i であるとき，提案分布を以下のように定める．

$$
q_c(c|c_i) = \mathrm{Gamma}(c|\gamma c_i, \gamma) \qquad (20)
$$

$$
q_d(d|d_i) = \mathrm{Gamma}(d|\delta d_i, \delta) \qquad (21)
$$

ここで，γ と δ は提案分布のハイパーパラメータである．$q_c(c|c_i)$ からサンプルされた c^* を用いて，以下のように採択率を計算する．

$$
g_c(c^*, c_i) = \min\left\{1, \frac{f_c(c^*)q_c(c_i|c^*)}{f_c(c_i)q_c(c^*|c_i)}\right\} \qquad (22)
$$

ここで，$f_c(c)$ は以下のように計算される尤度関数である．

$$
\begin{aligned}
f_c(c) &\equiv p(c|x_{1:T}, z_{1:N}, \tau_{1:N}, \Theta, d_i) \\
&\propto \prod_{n=1}^{N} \rho_{\tau_n} b_{z_n\tau_{n-1}\tau_n} \prod_{n=2}^{N} a_{z_{n-1}z_n}\pi_{z_1} q(c) \\
&= \prod_{n=1}^{N} \rho_{\tau_n} \left\{ \prod_{t=\phi_{n-1}}^{\phi_n - 1} \mathrm{Cauchy}(x_t|\mu_{z_n}, \lambda_t^c) \right\}^{\frac{1}{\phi_n - \phi_{n-1}}} \\
&\quad \cdot \prod_{n=2}^{N} a_{z_{n-1}z_n}\pi_{z_1} \mathrm{Gamma}(c|c_0, c_1) \qquad (23) \\
\lambda_t^c &= c_i \cdot \Delta x_t + d_i \qquad (24)
\end{aligned}
$$

そして，もし $g_c(c^*, c_i)$ の値が区間 $[0, 1]$ における一様分布からサンプルされた乱数 r よりも大きければ，$c_{i+1} = c^*$ とし，そうでなければ，$c_{i+1} = c_i$ とする．ただし，c_0 は事前分布 $q(c)$ からサンプルされる．

d の値は c と同様の方法で更新される．$q_d(d|d_i)$ からサンプルされた d^* を用いて，以下のように採択率を計算する．

$$
g_d(d^*, d_i) = \min\left\{1, \frac{f_d(d^*)q_d(d_i|d^*)}{f_d(d_i)q_d(d^*|d_i)}\right\} \qquad (25)
$$

ここで，$f_d(d)$ は以下のように計算される尤度関数である．

$$
\begin{aligned}
f_d(d) &\equiv p(d|x_{1:T}, z_{1:N}, \tau_{1:N}, \Theta, c_{i+1}) \\
&\propto \prod_{n=1}^{N} \rho_{\tau_n} b_{z_n\tau_{n-1}\tau_n} \prod_{n=2}^{N} a_{z_{n-1}z_n}\pi_{z_1} q(d) \\
&= \prod_{n=1}^{N} \rho_{\tau_n} \left\{ \prod_{t=\phi_{n-1}}^{\phi_n - 1} \mathrm{Cauchy}(x_t|\mu_{z_n}, \lambda_t^d) \right\}^{\frac{1}{\phi_n - \phi_{n-1}}} \\
&\quad \cdot \prod_{n=2}^{N} a_{z_{n-1}z_n}\pi_{z_1} \mathrm{Gamma}(d|d_0, d_1) \qquad (26) \\
\lambda_t^d &= c_{i+1} \cdot \Delta x_t + d_i \qquad (27)
\end{aligned}
$$

そして，もし $g_d(d^*, d_i)$ の値が区間 $[0, 1]$ における一様分布からサンプルされた乱数 r よりも大きければ，$d_{i+1} = d^*$ とし，そうでなければ $d_{i+1} = d_i$ とする．ただし，d_0 は事前分布 $q(d)$ からサンプルされる．

3.5 ビタビ復号

音符を表す潜在変数系列は学習過程において以下の式 (28) で与えられる尤度が最大の時のパラメータを用いたビタビアルゴリズムによって推定される．

表 1: 平均一致率と標準誤差

モデル	一致率
SBS-SHMM	66.3 ± 1.0
多数決法	56.9 ± 1.1
フレームベース HMM	56.1 ± 1.1
BS-HMM	67.0 ± 1.0

$$p(x_{1:T}) = \sum_{z_N=1}^{K} \sum_{\tau_N=-G}^{G} \alpha_{z_N \tau_N} \tag{28}$$

推定したい音符は $p(\boldsymbol{Z}, \boldsymbol{\tau}|\boldsymbol{X})$ を最大化する潜在変数の値である．ビタビアルゴリズムにおいて $\omega_{z_n \tau_n}$ を以下のように定める．

$$\omega_{z_n \tau_n} = \max_{\substack{z_{1:n-1} \\ \tau_{1:n-1}}} \ln p(X_{0\tau_1}, \ldots, X_{\tau_{n-1}\tau_n}, z_{1:n-1}, z_n, \tau_{1:n-1}, \tau_n) \tag{29}$$

そして，$\omega_{z_n \tau_n}$ は以下のように再帰的に計算される．

$$\omega_{z_1 \tau_1} = \ln \rho_{\tau_1} + \ln b_{z_1 0 \tau_1} + \ln \pi_{z_1} \tag{30}$$

$$\omega_{z_n \tau_n} = \ln \rho_{\tau_n} + \max_{\tau_{n-1}} \left[\ln b_{z_n \tau_{n-1} \tau_n} + \max_{z_{n-1}} \left\{ \ln a_{z_{n-1} z_n} + \omega_{z_{n-1} \tau_{n-1}} \right\} \right] \tag{31}$$

$\omega_{z_n \tau_n}$ の再帰計算において，$(z_n, \tau_n)=(k, g)$ の時の $\omega_{z_n \tau_n}$ の値を最大化する状態が $(z_{n-1}, \tau_{n-1})=(j, f)$ である場合，これらの状態は $h^{(z)}_{nk} = j, h^{(\tau)}_{ng} = f$ として記録される．$\{\omega_{z_N \tau_N}\}^{K,G}_{z_N=1, \tau_N=-G}$ が計算された後，式 (30) と (31) を用いて，潜在変数系列 $\{z_n, \tau_n\}^N_{n=1}$ が以下のように再帰的に推定される．

$$(z_N, \tau_N) = \arg\max_{z_N, \tau_N} \{\omega_{z_N \tau_N}\} \tag{32}$$

$$z_n = h^{(z)}_{(n+1) z_{n+1}} \tag{33}$$

$$\tau_n = h^{(\tau)}_{(n+1) \tau_{n+1}} \tag{34}$$

推定された潜在変数 $\{\tau_n\}^N_{n=1}$ によって表されるオンセット変動を補正することにより音符系列が得られる．

4. 評価実験

提案法と従来法を用いて歌声 F0 軌跡から推定された音符列の精度を評価する実験を行った．

4.1 実験条件

実験には RWC データベース [22] 内のポピュラー音楽 100 曲を用いた．それぞれの曲に対して，モデルパラメータの学習，音符系列の推定，推定された音符系列の精度の計測を行った．入力の F0 軌跡は池宮ら [2] の手法を用いてモノラル音楽音響信号から得る．また，ビート時刻は Durand ら [23] のビートトラッキングシステムによって得られたものを用いる．このシステムは全音符単位のビート時刻を推定するので，推定されたビート区間を 16 等分することで 16 分音符単位のビート時刻を求めた．

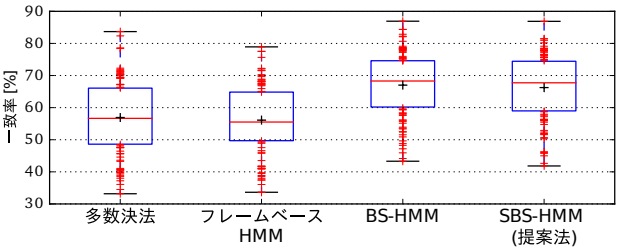


図 3: 一致率 [%]. 箱ひげ図内で，赤線は中央値，青色の箱は第 1 四分位から第 3 四分位の範囲，黒色の十字は平均値，赤色の十字は外れ値を示す．

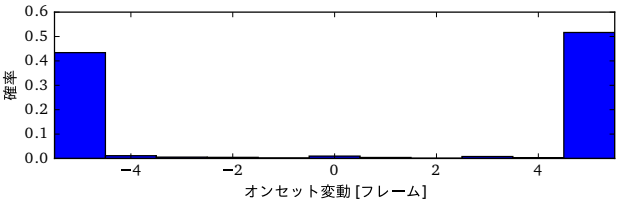


図 4: 学習されたモデルパラメータ $\boldsymbol{\rho}$ の例．

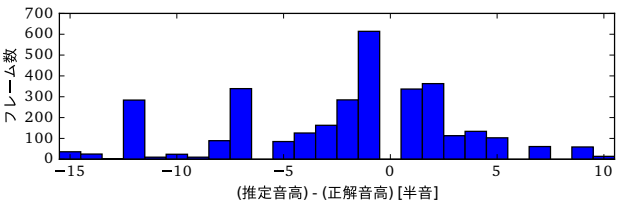


図 5: 音高推定誤りの例．推定音高が正しい場合は省略．

提案法のハイパーパラメータは $\boldsymbol{\xi} = \mathbb{1}$，$\boldsymbol{\zeta} = \mathbf{1}$，$\boldsymbol{\eta} = \mathbf{1}$，$c_0 = d_1 = d_0 = d_1 = 1$ とした．ここで，$\mathbb{1}$ と $\mathbf{1}$ はそれぞれ全要素が 1 である行列とベクトルである．提案分布のパラメータは $\gamma = \delta = 1$ とした．τ_n が取りうる最大値 G は 5（つまり 50 ミリ秒）とした．

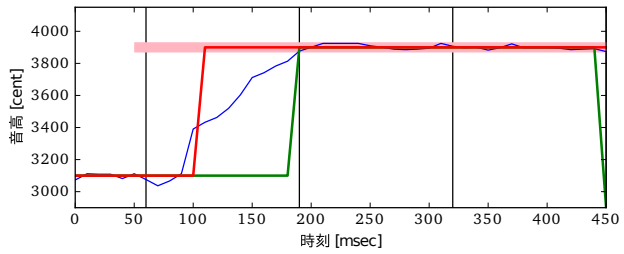
ベースラインとして多数決法をテストした．16 部音符に対応する時間区間ごとに歌声 F0 の多数決をとることで音高を決定する．比較として，フレームベース HMM とビート同期 HMM（BS-HMM）もテストした．フレームベース HMM は全てのビート区間が 1 フレームのみであるとする．BS-HMM はオンセット変動を考慮しないこと以外 SBS-SHMM と同じである．

推定された音符列は楽曲のメロディと同期した MIDI データと比較し，一致率（つまり，音高が正しく推定されているフレームの比率）を評価尺度として用いた．

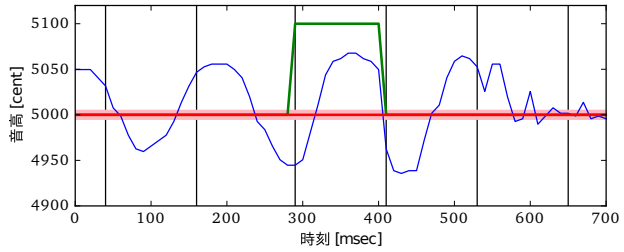
4.2 実験結果

音符推定の結果を表 1 と図 3 に示した．提案法は平均一致率において多数決法とフレームベース HMM を大きく上回った．一方で，提案法と BS-HMM の一致率はほぼ等しく，その差は統計的に有為なものではなかった．

この結果は遷移確率による楽譜モデルと出力確率による周波数変動モデルが音符推定精度の向上に対して有効であることを



(a) オンセット変動



(b) 周波数変動

図 6: 音高推定結果の例．桃色，青色，緑色，赤色，黒色の線はそれぞれ MIDI ノート，オンセット変動を含む F0 軌跡，多数決法によって推定された音高，提案法によって推定された音高，事前に推定されたビート時刻を表す．

示している．オンセット変動のモデルにより精度が向上しなかったのは，モデルパラメータ $\boldsymbol{\rho}$ が正しく学習されなかったことが原因だと考えられる（図 4）．オンセット変動はオンセットの両側の音高の長さや曲全体のテンポに依存するので，単一の離散分布でオンセット変動を捉えるのは難しい．音高が遷移している間の F0 を表現するための隠れ状態を用いるなどして，より詳細にオンセット変動をモデル化することが必要だろう．

4.2.1 音高推定誤り

主に二種類の誤りが観測された（図 5）．一つ目は歌手の歌唱表現によるものであり，1 半音や 2 半音の誤りとして現れる．これは周波数変動が音符推定精度に影響することを意味する．二つ目は F0 推定誤りによるものであり，7 半音や 12 半音誤りとして現れる．7 半音や 12 半音は完全五度や 1 オクターブに対応する．

4.2.2 歌唱表現抽出と頑健性

図 6 の音符推定結果の例は歌手の歌唱表現を提案モデルが捉える様子を示している．上の図において，最初のビートでのオンセットはビート時刻より遅れている．提案法ではオンセットの遅れを正しく捉えているのに対し，多数決法はオンセットとすべきビート時刻を誤認識している．下の図はビブラートの例である．多数決法では推定結果が大きな周波数変動に影響されている．一方，提案法ではコーシー分布の頑健性によりビブラートに影響されることなく正しく音高が推定できている．

5. おわりに

本稿では，ビート時刻を既知として歌声 F0 軌跡から楽曲の音符推定を行う手法を提案した．歌声 F0 軌跡の生成過程をモデル化するに際し，楽譜成分だけでなくオンセット・周波数変動を考慮した．SBS-SHMM は多数決法やフレームベース HMM よりも正確な音高推定を実現した．

提案法を用いて得られたオンセット変動と周波数変動は歌唱表現の特徴を捉えるうえで重要である．今後は，2 次系伝達関数などを用いて歌声 F0 軌跡を詳細にモデル化し，歌唱表現を種類ごとに抽出する手法の開発を行いたい．提案法では，F0 推定，ビートトラッキング，音符推定は別々に行われたが，これらの手法を統合することも必要である．また，実際の楽曲に存在する無声部分を扱うことができなかったが無声部分を適切に扱えるようにしたい．

謝辞：本研究の一部は JST OngaCREST プロジェクト，JSPS 科研費 24220006，26700020，26280089，15K16054，16H01744，16J05486 と栢森情報科学振興財団助成金による支援を受けて行われた．

文　献

[1] D.J. Hermes, “Measurement of pitch by subharmonic summation,” The journal of the acoustical society of America, vol.83, no.1, pp.257–264, 1988.

[2] Y. Ikemiya, K. Yoshii, and K. Itoyama, “Singing voice analysis and editing based on mutually dependent F0 estimation and source separation,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.574–578, 2015.

[3] M. Goto, “A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals,” Speech Communication, vol.43, no.4, pp.311–329, 2004.

[4] J. Salamon and E. Gómez, “Melody extraction from polyphonic music signals using pitch contour characteristics,” IEEE Transactions on Audio, Speech, and Language Processing, vol.20, no.6, pp.1759–1770, 2012.

[5] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, “Source/filter model for unsupervised main melody extraction from polyphonic audio signals,” IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.3, pp.564–575, 2010.

[6] A. deCheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” The Journal of the Acoustical Society of America, vol.111, no.4, pp.1917–1930, 2002.

[7] M. Mauch and S. Dixon, “pYIN: A fundamental frequency estimator using probabilistic threshold distributions,” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.659–663, 2014.

[8] Y. Li and D. Wang, “Separation of singing voice from music accompaniment for monaural recordings,” IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.4, pp.1475–1487, 2007.

[9] P.-S. Huang, S.D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, “Singing-voice separation from monaural recordings using robust principal component analysis,” 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.57–60, 2012.

[10] Y.E. Kim and B. Whitman, “Singer identification in popular music recordings using voice coding features,” The 3rd International Conference on Music Information Retrieval, pp.164–169, 2002.

[11] W.-H. Tsai and H.-M. Wang, “Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals,” IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.1, pp.330–341, 2006.

[12] A. Dobashi, Y. Ikemiya, K. Itoyama, and K. Yoshii, “A mu-

sic performance assistance system based on vocal, harmonic, and percussive source separation and content visualization for music audio signals," The 12th Sound and Music Computing Conference (SMC), pp.99–104, 2015.
[13] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," 2008 IEEE International Conference on Multimedia and Expo (ICME), pp.1417–1420, 2008.
[14] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, "Songle: A web service for active music listening improved by user contributions," The 12th International Society for Music Information Retrieval Conference (ISMIR), pp.311–316, 2011.
[15] R.P. Paiva, T. Mendes, and A. Cardoso, "On the detection of melody notes in polyphonic audio," The 6th International Conference on Music Information Retrieval (ISMIR), pp.175–182, 2005.
[16] C. Raphael, "A graphical model for recognizing sung melodies," The 6th International Conference on Music Information Retrieval (ISMIR), pp.658–663, 2005.
[17] G.E. Poliner and D.P.W. Ellis, "A classification approach to melody transcription," The 6th International Conference on Music Information Retrieval (ISMIR), pp.161–166, 2005.
[18] A. Laaksonen, "Automatic melody transcription based on chord transcription," The 15th International Society for Music Information Retrieval Conference (ISMIR), pp.119–124, 2014.
[19] M.P. Ryynänen and A.P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," Computer Music Journal, vol.32, no.3, pp.72–86, 2008.
[20] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the Tony software: Accuracy and efficiency," The First International Conference on Technologies for Music Notation and Representation (TENOR), pp.23–30, Institut de Recherche en Musicologie, Paris, France, 2015.
[21] Y. Ohishi, H. Kameoka, D. Mochihashi, and K. Kashino, "A stochastic model of singing voice F0 contours for characterizing expressive dynamic components," The 13th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.474–477, 2012.
[22] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," The 3rd International Conference on Music Information Retrieval (ISMIR), vol.2, pp.287–288, 2002.
[23] S. Durand, J.P. Bello, B. David, and G. Richard, "Downbeat tracking with multiple features and deep neural networks," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.409–413, 2015.