3Q-05

ビート準同期隠れマルコフモデルに基づく歌声音高軌跡に対する音符推定

錦見 亮 † 中村 栄太 ‡ 糸山 克寿 ‡ 吉井 和佳 ‡

† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

歌声の音高 (F0) 軌跡を適切な数理モデルにより表現することは，複雑な音楽音響信号中に含まれる歌声の F0 軌跡を推定するうえで有用である．例えば，音楽鑑賞支援サービス Songle [1] では，Web 上の任意の音楽音響信号中の歌声の F0 軌跡を推定し可視化する．このとき，人間の歌声として自然な F0 軌跡とはこういうものであるという事前知識があれば，より正確な F0 推定が可能になるであろう．また，推定された F0 軌跡から楽譜上の音符系列 (メロディ) に基づく成分と歌手の個性に基づく成分を分離することは，メロディが似ている楽曲の検索や歌唱表現が似ている歌手の検索など [2] の実現に有用である．

混合音中の歌声 F0 軌跡を推定する研究は従来より盛んに行われてきたが [3]，離散的な音符系列に変換する研究はごくわずかである．歌声 F0 軌跡を音符系列に変換するもっとも単純な方法は，Songle に実装されているように，まず F0 軌跡を半音単位でクオンタイズし，最小の時間単位 (例えば 16 分音符単位) ごとに多数決を取る方法である．しかし，ビブラートやオーバーシュートといった大幅な F0 変動を伴う歌唱表現や，正確なビート時刻からの進みや遅れの影響で，変換精度には限界がある．大石らは歌声 F0 軌跡の生成過程を隠れマルコフモデル (HMM) を用いてモデル化し，階段状に変化する楽譜成分と局所的に変化する逸脱成分とに分離する手法を提案している [4]．ただし，楽譜上の音符系列の生成過程はモデル化されておらず，音符推定は扱われていなかった．Raphael は HMM に基づき，独唱歌声音響信号からの音符推定を行っている [5]．

本稿では，歌声 F0 軌跡に対する音符推定のためのビート準同期 HMM を提案する (図 1)．本モデルは，楽譜上の音高が厳密にビートに同期して遷移することで階段状の F0 軌跡が生成され，さらに周波数方向・時間方向ともに変動成分 (ずれ) が生じることで実際の歌声 F0 が生成される過程を定式化したものである．我々は，歌声 F0 軌跡が観測データとして与えられた時に，潜在変数である楽譜上の音符系列および実際の発音時刻とビート時刻とのずれを推定する問題を解く．ただし，本研究ではビート時刻は既知であることを仮定しているので，予め [6, 7] 等を用いてビート時刻を推定しておく必要がある．

2. 提案手法

本手法の入力歌声 F0 軌跡は一定時間ごとにサンプルした対数周波数 $\{x[t]\}_{t=1}^{T}$ として表される．ここで，t はフレーム番号であり，T は入力信号の長さを表す．また，n 番目のビート時刻を $\psi[n] \in \{1, \ldots, T+1\}$ と表す．曲の始端は $\psi[0] = 1$，終端は $\psi[N] = T+1$ と表される．F0 軌跡の楽譜成分は，$\psi[n-1]$ から $\psi[n]$ までの各

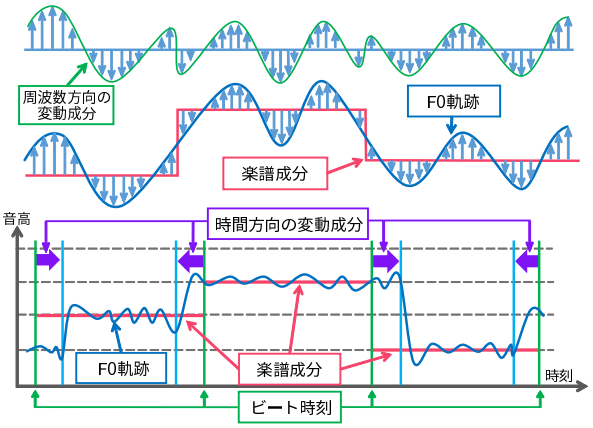

図 1: モデルの全体像

ビート区間に，1 つずつ割り当てられた半音単位の音高 $z[n] = 1, \ldots, K$ の系列により表される (K は楽譜に現れうる半音の数)．各ビート n に対する歌声の音高変化時刻を $\phi[n]$ と記し，歌声の音高変化時刻とビート時刻とのずれ $\tau[n] := \phi[n] - \psi[n]$ は $-G$ から G までの値を取りうるものとする．以下，$\phi[n-1]$ から $\phi[n]$ までの区間を伸縮ビート区間と呼ぶ．

歌声 F0 軌跡の生成モデルを構成するために，$\{\tau[n], z[n]\}_{n=1}^{N}$ と $\{x[t]\}_{t=1}^{T}$ は確率的に生成されるものとする．$\tau[n]$ は各ビート n に関して，独立な離散分布に従い生成されるものとする．隣接音高間の統計的依存性を表現するため，$z[n]$ の生成確率は直前の音高 $z[n-1]$ に依存する離散分布に従うものとする．n 番目の伸縮ビート区間に含まれる各フレームでの観測 F0 $x[t]$ の値は，音高 $z[n] = k$ の対数周波数 μ_k と，確率的に値をとる歌声の周波数方向の逸脱成分との和で生成されるものとする．各フレームでの $x[t]$ の生成確率が独立であると仮定すると，以上のモデルは HMM として記述できる．

歌声 F0 軌跡は逸脱成分による多くの変動が含まれるため，各時刻 t についての出力確率には正規分布よりも外れ値に対して頑健なコーシー分布 $\mathrm{Cauchy}(x; \mu, \lambda) = \lambda / [\pi \{(x-\mu)^2 + \lambda^2\}]$ で記述する．μ は分布の最頻値を与える位置パラメータで，対応する音符の音高を $z[n] = k$ とすると $\mu = \mu_k$ となる．また，λ は半値半幅を与える尺度パラメータである．歌声にはグリッサンドのような変化量の大きな逸脱成分や，微細変動のような変化量の小さな逸脱成分が存在する．そこで，ΔF0 の絶対値 $|x[t] - x[t-1]|$ がある閾値 θ 以上か否かにより異なる尺度パラメータ λ_+ と λ_- を用いることで，音高変化量の異なる逸脱成分を表現して周波数方向のゆらぎを表現する．

以上のモデルを Bayesian HMM に基づいて定式化する．$z[n]$ の遷移確率行列を $\boldsymbol{A} = (a_{jk})_{j,k=1}^{K}$ と記すと，

$$p(z[n] = k | z[n-1] = j) = a_{jk} \quad (1)$$

$$\boldsymbol{a}_j \sim \mathrm{Dirichlet}(\boldsymbol{\xi}_j) \quad (2)$$

Automatic Note Estimation for Vocal F0s based on a Beat-Semi-Synchronous Hidden Markov Model: *Ryo Nishikimi, Eita Nakamura, Katsutoshi Itoyama, and Kazuyoshi Yoshii (Kyoto Univ.)*

表 1: 音符の音高推定の精度 [%](100 曲の平均 ± 分散)

多数決法	時間方向の ゆらぎなし	時間方向の ゆらぎあり
58.32 ± 12.00	66.76 ± 10.52	66.04 ± 10.28

と表される．ここで，$\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jK})$ であり，$\boldsymbol{\xi}_j$ は $\boldsymbol{a}_j$ が従うディリクレ分布のハイパーパラメータである．同様に，初期確率は

$$p(z[1] = k) = \pi_k, \quad \boldsymbol{\pi} \sim \mathrm{Dirichlet}(\boldsymbol{\zeta}) \tag{3}$$

とする．$\tau[n]$ の生成確率は

$$p(\tau[n] = g) = \rho_g, \quad \boldsymbol{\rho} \sim \mathrm{Dirichlet}(\boldsymbol{\eta}) \tag{4}$$

と表される．ここで，$\boldsymbol{\rho} = (\rho_{-G}, \cdots, \rho_G)$ であり，$\boldsymbol{\eta}$ は $\boldsymbol{\rho}$ が従う分布のハイパーパラメータである．伸縮ビート区間 $\phi[n-1] \leq t < \phi[n]$ に含まれる F0 値 $(x[t])_{t=\phi[n-1]}^{\phi[n]-1}$ の出力確率は以下で与えられる．

$$\begin{aligned} &p\big((x[t])_{t=\phi[n-1]}^{\phi[n]-1} | z[n] = k\big) \\ &= \left\{ \prod_{t=\phi[n-1]}^{\phi[n]-1} \mathrm{Cauchy}(x[t] | \mu_k, \lambda[t]) \right\}^{1/\Delta[n]} \\ &\Delta[n] \coloneqq \phi[n] - \phi[n-1] \\ &\lambda[t] = \begin{cases} \lambda_+ & (|x[t] - x[t-1]| \geq \theta) \\ \lambda_- & (|x[t] - x[t-1]| < \theta) \end{cases} \end{aligned} \tag{5}$$

モデルパラメータはギブスサンプリング法を用いて，潜在状態系列 $\{z[n], \tau[n]\}_{n=1}^N$ とパラメータ $\Theta = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\rho}\}$ を交互にサンプリングを行い学習する．また，出力確率のコーシー分布のパラメータ λ_+ と λ_- は，Θ の学習の時にサンプリングされた潜在状態系列を用いて [8] の手法により学習する．

3. 評価実験

提案法による歌声 F0 軌跡の楽譜成分推定に対する有効性を評価するため，RWC データベースのポピュラー音楽 100 曲 [9] を用いて評価した．入力の歌声 F0 にはモノラル音響音楽信号から池宮らの手法 [3] により推定されたものを用いた．ビート時刻はデータベース内の正解データ [10] を用いた．これには，4 分音符単位でのビート時刻が示されているが，楽譜上の音符の最小単位としては 16 分音符が適切であるため，4 分音符単位のビート時間を均等に 4 分割することで 16 分音符単位のビート時刻を求めた．音符系列はビタビ探索によって得られる伸縮ビート区間の潜在状態系列 $\{(z[n], \tau[n])\}_{n=1}^N$ から推定した．そして，時刻ずれを補正してビート区間に対して音高を割り当てた音符系列と，データベース内の同期 MIDI とをフレーム時刻単位 (10 ms) で比べ一致率を評価した．なお，学習過程においてハイパーパラメータは $\boldsymbol{\xi} = \mathbb{1}, \boldsymbol{\zeta} = \mathbf{1}, \boldsymbol{\eta} = \mathbf{1}$ とし，閾値は $\theta = 20$ cent とした．ここで $\mathbb{1}$ と $\mathbf{1}$ は，要素すべてが 1 の行列とベクトルを表す．

提案手法を多数決法と比較した．また，時間方向のゆらぎをモデル化することの有効性も評価するため，時間方向のゆらぎのないビート時刻でのみ音高遷移を許したモデルとも比較した．

評価実験の結果を表 1 に示す．結果から提案手法では

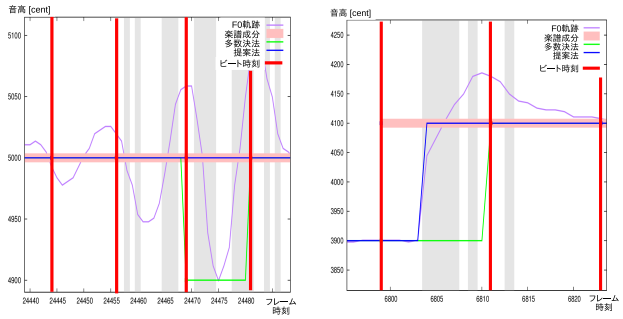

図 2: 音符推定結果の例．図中，グレーの背景になっているフレームは ΔF0 が閾値よりも大きかった場所を示している．

多数決法よりも音符列の推定精度が向上したことが分かる．ただし，時間方向のずれをモデル化することによる推定精度の有意な差異は見られなかった．ビート準同期 HMM を用いた音高推定結果の例を図 2 に示す．図 2 の左図では，変動が大きい歌声 F0 軌跡に対しても適切な音高が推定できている．また，図 2 の右図では，歌声 F0 軌跡の音高変化がビート時刻よりも遅れて起きているが，この歌声 F0 の音高変化の時刻ずれを自動的に抽出できていることが分かる．

4. 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．楽譜に指定された F0 軌跡からの周波数的・時間的な変動成分を生成モデルで記述することにより，従来法よりも優れた音高推定精度を達成した．本手法により得られた歌声 F0 の音高変化時刻のビート時刻からのずれや周波数的なずれは，歌手の歌唱表現の特徴を捉える上で重要であると考えられる．今後は 2 次系伝達関数などを用いて歌声 F0 軌跡の 周波数方向のずれを詳細にモデル化することにより，歌唱表現の種類ごとに逸脱成分を抽出する手法の開発を行いたい．また，本稿ではビート時刻を既知としていたが，ビートトラッキング手法 [6, 7] と本手法を統合することも今後の課題である．

謝辞 本研究の一部は，JSPS 科研費 24220006, 26700020, 26280089, JST CREST の支援を受けた．

参考文献

[1] 後藤真孝ほか: “Songle: ユーザが誤り訂正により貢献可能な能動的音楽鑑賞サービス,” 情報処理学会 インタラクション 2012 論文集, 1–8, 2012.

[2] J. A. Downie: “Music Information Retrieval,” *Ann. Rev. Inf. Sci. Tech.*, vol. 37, 295–340, 2003.

[3] 池宮由楽ほか: “音楽音響信号に対する相補的な歌声分離と音高推定,” 第 77 回情処全大, 5S-01, 2015.

[4] 大石康智ほか: “ノート指令と表現指令によって駆動される歌声 F0 生成過程の統計モデル,” 音講論集 (春), 2-11-6, 345–348, 2012.

[5] C. Raphael: “A Graphical Model for Recognizing Sung Melodies,” *ISMIR*, 658–663, 2005.

[6] G. Peeters: “Template-Based Estimation of Time-Varying Tempo,” *EURASIP J. Appl. Sig. Proc.*, 158–158, 2007.

[7] M. Khadkevich *et al.*: “A Probabilistic Approach to Simultaneous Extraction of Beats and Downbeats,” *ICASSP*, 445–448, 2012.

[8] J. H. MuCulloch: “Simple Consistent Estimators of Stable Distribution Parameters,” *J. Communications in Statistics-Simulation and Computation*, 1109-1136, 1986.

[9] M. Goto *et al.*: “RWC Music Database: Popular, Classic, and Jazz Music Databases,” *ISMIR*, 287–288, 2002.

[10] M. Goto: “AIST Annotation for the RWC Music Database,” *ISMIR*, 359–360, 2006.