7L-01

スケールと音高の過渡的遷移を考慮したHSMMに基づく歌声F0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 大学院情報学研究科 知能情報学専攻

1. はじめに

歌声は通常ポピュラー音楽のメロディを担っており，楽曲の印象に大きな影響を与えたり歌手の個人性を反映していたりするため，歌声解析は音楽情報処理において重要なテーマの1つである．楽曲に対する歌声基本周波数（F0）推定 [1,2] や歌声分離 [1,3] が広く研究されており，これらの技術は歌手同定 [4] やカラオケシステム [5]，能動的音楽鑑賞システム [6] などに用いられている．

本研究では，歌声 F0 軌跡に対する音符推定問題に取り組む．従来，この問題に対する様々な手法が提案されている．能動的音楽鑑賞システム Songle [6] に実装されている多数決法では，歌声 F0 軌跡を一番近い半音に離散化し，一定時間単位（例えば 16 分音符単位）ごとに多数決で音高を決定する．Laaksonen [7] は事前に与えたコード情報から楽曲の調と音符のセグメントを推定した後，調，コード，楽曲に依存するスコア関数を用いて各セグメントに対する音符の音高を決定する手法を提案している．Ryynänen ら [8] は音符内の状態遷移を left-to-right な隠れマルコフモデル（HMM）で表現し，音符間の音高遷移を事前に推定したキーに依存させることでメロディとベースラインを推定する手法を提案している．

本稿では，歌声 F0 軌跡とビート時刻を入力として，楽曲のメロディの音符系列を確率モデルに基づいて推定する手法を提案する（図 1）．従来法 [9] では，ビート時刻と歌声の音高変化開始時刻とのずれ（オンセット変動）や音符の音高と歌声の音高とのずれ（周波数変動）が楽譜に付加されることで歌声 F0 軌跡が生成される過程を HMM を用いてモデル化している．本手法では，より自然な楽譜が推定されることを目指して，楽譜が調やリズムに依存して生成される過程を従来法のモデルに組み込む．また，従来法では歌声における時間方向の変動に関してオンセット変動しか考慮されていない．本手法では，実際の歌声の音高がある程度の時間を要して徐々に変化することを捉えるため，歌声の音高変化に要する時間を表す潜在変数を導入する．

2. 提案手法

本章では，調に依存して楽譜が生成され，その楽譜に時間・周波数方向の変動が付加されて歌声が生成される過程を表現する手法について述べる（図 1）．本手法の入力歌声 F0 軌跡は対数周波数（単位は cent）の系列 $X = \{x_t\}_{t=1}^{T}$ で表される．ここで，t はフレーム番号，T は歌声 F0 軌跡のフレーム数を表す．また，n 番目のビート時刻（16 分音符単位）を ψ_n とし，曲の始端は ψ_0=1，終端は ψ_N=T+1 で表す．

調の遷移モデルについて説明する．調 $Y = \{y_m\}_{m=1}^{M}$（M は曲中の小節数）は各小節ごとに割り当てられる．調を 1 曲に対して 1 つに固定するのではなく，小節単位

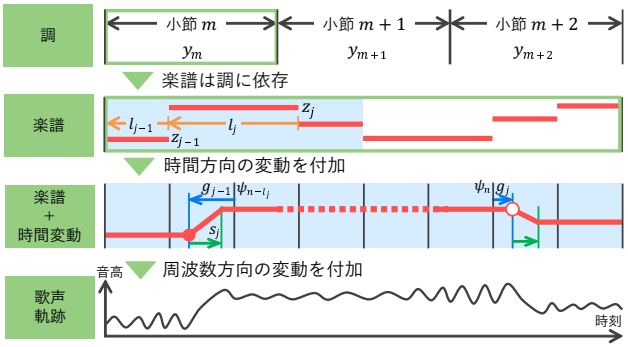

図 1: モデルの全体像

での遷移を許すことで，転調に対応できるようにする．第 m 小節の状態 y_m は $\{\mathrm{C}, \mathrm{C\#}, \ldots, \mathrm{B}\} \times \{\mathrm{Major}, \mathrm{Minor}\}$ の 24 通りの値を取りうる．調 Y はマルコフモデルに従い，

$$p(y_1) = v_{y_1}, \quad p(y_m|y_{m-1}) = u_{y_{m-1}y_m} \tag{1}$$

である．ここで，v_{y_1} と $u_{y_{m-1}y_m}$ はそれぞれ調の初期確率と遷移確率を表す．

調に依存して楽譜中の音符列が生成される過程について説明する．楽譜中の音符列を $Z = \{(m_j, z_j, h_j)\}_{j=1}^{J}$（$J$ は楽譜中の音符数）と記す．ここで，$m_j \in \{1, \ldots, M\}$ は音符 j のオンセットが属する小節番号，$z_j \in \{1, \ldots, K\}$ は音符の音高（半音単位），K は楽譜に現れうる半音の数，$h_j \in \{1, \ldots, 16\}$ は音符 j のオンセットのビート位置（16 分音符単位）を表す．また，小節 m 内にオンセットが属している音符の中で一番最初の音符を $j(m)$ と記す．音符 j の音高の生成確率は，一つ前の音高と間の遷移確率 $a_{z_{j-1}z_j} = p(z_j|z_{j-1})$ と音符のオンセットが属する小節 m の調に依存するピッチクラスの生成確率 $w_{y_m\tilde{z}_j} = p(\tilde{z}_j|y_m)$ の重み付き積で記述する．ここで $\tilde{z}_j (\equiv z_j \bmod 12)$ は z_j のピッチクラスである．重み係数を κ とする時，小節 m 内の音高の生成確率は以下で与えられる

$$p(\{z_j\}_{j=j(m)}^{j(m+1)-1}|y_m) \propto \prod_{j=j(m)}^{j(m+1)-1} a_{z_{j-1}z_j}^{\kappa} w_{y_m\tilde{z}_j}^{1-\kappa}.$$

音符 j のオンセット位置 h_j および長さ l_j=h_{j+1}−h_j は，1 つ前の音符のオンセット位置に依存するマルコフモデルに従って生成されるものとする．オンセット位置の遷移確率を以下の通り記す

$$p(h_j|h_{j-1}) = d_{h_{j-1}h_j}. \tag{2}$$

楽譜 Z から歌声 X が生成される過程のモデル化について説明する．歌声は楽譜に時間方向の変動と周波数方向の変動が付加されることで生成される．時間方向の変動は 2 つの変数 (d_j, s_j) で表す．ここで，$g_j \in \{-G, \ldots, G\}$ はビート時刻と歌声の音高遷移開始時刻とのずれ（フレーム単位），$s_j \in \{1, \ldots, S\}$ は音高の遷移にかかる時間（フレーム単位）を表す．d_j と s_j は各音符において以下

Musical Note Estimation from Vocal F0 Trajectories Based on HSMM Considering Scales and Transitional Change of Pitches. *Ryo Nishikimi, Eita Nakamura, Katsutoshi Itoyama, and Kazuyoshi Yoshii (Kyoto Univ.)*

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.

表 1: 音高推定精度 [%]（平均一致率と標準偏差）

多数決法	調依存性なし	調依存性あり
57.04 ± 11.48	60.06 ± 10.56	61.33 ± 10.16

の通り独立に離散分布に従って生成される

$$p(g_j) = b_{g_j}, \quad p(s_j) = c_{s_j}. \tag{3}$$

周波数方向の変動の付加は，時間変動が付加された音符の音高から実際の歌声の音高が確率的に生成される過程を表す．音符 j のオンセットの時刻がビート時刻 ψ_n である時，音符 j から歌声が生成される確率は，時間方向の変動を付加した音符内の各フレーム $(\tau_j^0{=}\psi_n{+}g_{j-1}{\leq}\, t\, {<}\tau_j^1{=}\psi_{n+l_j}{+}g_j)$ における対数周波数 x_t の出力確率ので定義する．歌声 x_t の出力確率は正規分布とすると，音符 j からの歌声の出力確率は

$$p(\{x_t\}_{t=\tau_j^0}^{\tau_j^1-1}|Q_j, Q_{j-1}) = \prod_{t=\tau_j^0}^{\tau_j^1-1} \mathcal{N}(x_t|\mu_t, \sigma^2) \tag{4}$$

で定義する．ここで，$Q_j{=}(z_j, d_j, s_j, l_j)$ であり，σ^2 は正規分布の分散パラメータである．また，μ_t は正規分布の平均パラメータで，以下のように定義する．

$$\mu_t{=}\begin{cases}(\mu_{z_j}{-}\mu_{z_{j-1}})/s_j{\cdot}(t{-}\tau_j^0){+}\mu_{z_{j-1}} & (\tau_j^0{\leq}t{<}\tau_j^0{+}s_j)\\ \mu_{z_j} & (\tau_j^0{+}s_j{\leq}t{<}\tau_j^1)\end{cases} \tag{5}$$

ここで，μ_{z_n} は音高 z_n に対応した対数周波数である．

提案手法をベイズ化するために，離散分布のパラメータ $u_*, v_*, w_*, a_*, b_*, c_*, d_*$ に対してはディリクレ共役事前分布を，正規分布の精度パラメータ $(\sigma^2)^{-1}$ に対してはガンマ共役事前分布を置く．モデルの学習は，まず，入力の F0 に対して多数決法を用いて音符系列を初期化する．次に，ギブスサンプリング法を用いてを順番に潜在変数とパラメータの事後分布を推論する．最後に，学習過程において潜在変数とパラメータの同時事後分布が最大になったときパラメータを用いてビタビ探索により最終的な音符系列を推定する．

3. 評価実験

提案法による音符推定精度を評価する．楽曲は RWC データベース [10] のポピュラー音楽 100 曲のうち，32 分音符や 3 連符を含む曲や，ボーカルパートに副旋律が含まれる曲などを除いた 73 曲を用いる．入力の歌声 F0 は池宮らの手法 [1] により推定されたものを，ビート時刻や小節情報はデータベース内のアノテーションデータ [11] を用いる．提案手法ではビートの最小単位に 16 分音符を仮定しているが，アノテーションは 4 分音符単位のビート時刻が記されているため，4 分割することで 16 分単位のビート時刻を得た．提案法を用いて推定された音符系列をデータベース内の MIDI データを基に作成した楽譜を 16 分音符単位で比較し，音高の一致率を評価する．学習過程においてハイパーパラメータは調遷移と音高遷移は自己遷移が高くなるように，調からピッチクラスが出力される確率は調に対応したスケール内の音が出やすくなるように設定した．

提案手法を多数決法と比較した．また，調依存性の有効性を評価するため，調に依存させない場合の音符推定精度も評価した．実験結果を表 1 に示す．結果から，多数決法よりも提案手法を用いた場合に音符推定精度が向上したことが分かる．また，調依存性なしの場合よりも調依存性ありの場合の方が精度が向上したことから，調依存性の有効性も示せた．提案手法による音符推定結果の例を図 2 に示す．左図においては，時間方向の変動を含む楽譜（灰色の点線）が音高変化が大きい歌声 F0 軌跡にフィットしており，正しく音符の音高が推定できている．また右図においては，ビブラートのような激しい音高の変化に対しても，多数決法で半音誤っているが，提案法では正しい音高が推定できている．

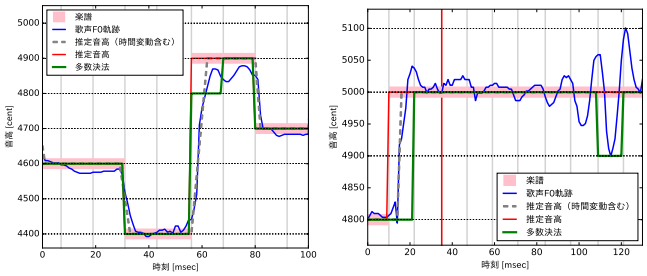

図 2: 音符推定結果の例

4. おわりに

本稿では，ビート時刻や小節情報を既知として歌声 F0 軌跡から楽曲のメロディの音符系列を推定する手法を提案した．言語モデルと音響モデルの統合モデルを用いて歌声 F0 軌跡の生成過程を表現することにより，従来法よりも音高推定精度が向上した．本手法を用いることにより，音符系列を推定するだけではなく，歌声 F0 軌跡に含まれる変動や楽曲の調を推定することができる．これらは，歌手の歌唱表現や楽曲の構造を捉える上で有用であると考えられる．今後は，歌声 F0 軌跡の周波数方向の変動に対する詳細なモデル化や VAD の導入によってさらなる音符推定精度の向上を目指す．

謝辞 本研究の一部は，JSPS 科研費 16J05486, 15K16054, 24220006, 26700020, 26280089, 16H01744, JST CREST，JST ACCEL の支援を受けた

参考文献

[1] Y. Ikemiya *et al.*: “Singing Voice Analysis and Editing based on Mutually Dependent F0 Estimation and Source Separation,” *ICASSP*, 574–578, 2015.

[2] J. -L. Durrieu *et al.*: “Source/filter Model for Unsupervised Main Melody Extraction from Polyphonic Audio Signals,” *IEEE*, 564–575, 2010.

[3] P. -S. Hung *et al.*: “Singing-voice Separation from Monaural Recordings Using Robust Principal Component Analysi,” *ICASSP*, 57–60, 2012.

[4] Y. E. Kim *et al.*: “Singer Identification in Popular Music Recordings Using Voice Coding Features,” *ISMIR*, 164–169, 2002.

[5] M. Ryynänen *et al.*: “Accompaniment Separation and Karaoke Application based on Automatic Melody Transcription,” *ICME*, 1417–1420, 2008.

[6] M. Goto *et al.*: “Songle: A Web Service for Active Music Listening Improved by User Contributions,” *ISMIR*, 311–316, 2011.

[7] A. Laaksonen *et al.*: “Automatic Melody Transcription based on Chord Transcription,” *ISMIR*, 119–124, 2014.

[8] M. P. Ryynänen *et al.*: “Automatic Transcription of Melody, Bass Line, and Chords in Polyphonic Music,” *Computer Music Journal*, 72–86, 2008.

[9] R. Nishikimi: *et al.* “Musical Note Estimation for F0 Trajectories of Singing Voices Based on a Bayesian Semi-beat-synchronous HMM,” *ISMIR*, 461–467, 2016.

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

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

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.