一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IBISML2017-54 (2017-11)

[ポスター講演] 調とリズムを考慮した階層隠れセミマルコフモデルに基づく歌声の自動採譜

錦見　亮†　中村　栄太†　後藤　真孝††　糸山　克寿†　吉井　和佳†,†††

† 京都大学大学院情報学研究科
†† 産業技術総合研究所 (AIST)
††† 理化学研究所 革新知能統合研究センター (AIP)
E-mail: †{nishikimi,enakamura,itoyama,yoshii}@sap.ist.i.kyoto-u.ac.jp, ††m.goto@aist.go.jp

あらまし　本稿では歌声 F0 軌跡から音楽的に自然な音符系列を推定する統計的手法を示す．歌声の発音時刻や F0 は楽譜に示されたビート時刻や音符の音高からの大きな逸脱を含むため，歌声 F0 軌跡の時間・周波数方向への離散化による音符推定の精度を向上するためには，楽譜の音楽的な自然さを表現する楽譜モデルが重要である．我々は調とリズムに依存する音符の音高を表現する楽譜モデルと楽譜 (音符系列) から時間・周波数方向に逸脱する歌声 F0 軌跡を表現する F0 モデルとを統合した階層隠れセミマルコフモデル (HHSMM: hierarchical hidden semi-Markov model) を提案する．楽譜モデルでは，確率的に生成された調に従って音符の音高が生成される．さらに，音符の開始位置はビートの 1 次元格子上に定義されたマルコフ過程に従って生成される．F0 モデルでは，歌声の発音時刻の時間方向の逸脱，音符間における F0 の滑らかな遷移，F0 の周波数方向の逸脱が確率的に生成され，楽譜に付与される．提案法では，楽譜モデルと F0 モデルが音符推定に与える影響を考慮しながら，入力の歌声 F0 軌跡から尤もらしい音符系列を推定する．実験結果から調やリズムを考慮しない場合と比較して，提案法による音符系列の推定精度が向上することを示した．

キーワード　歌声，自動採譜，階層隠れセミマルコフモデル

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，楽曲に関する多くの情報を提供するため，歌声解析は音楽情報検索をはじめとした様々な音楽アプリケーションにとって重要である．歌声 F0 推定 [1–5] や歌声分離 [6,7] といった歌声解析技術は盛んに研究されており，歌手同定 [8,9]，カラオケ生成 [10]，ハミング検索 [7]，能動的音楽鑑賞 [11] などに応用されている．さらに，歌声に含まれる情報をより活用するためには，歌声 F0 軌跡を離散的な記号のみを含む楽譜に変換することが有用である．

本研究では，楽譜に対して多くの逸脱を含む歌声 F0 軌跡からの音符系列推定に取り組む．楽譜における音符の音高や開始位置は離散的な値であるが，歌声 F0 軌跡は時間経過ともに滑らかに変化する連続的な信号である．例えば，歌声 F0 軌跡はビブラートによって振動したり，ポルタメントによってある音符から次の音符へと滑らかに変化したりする．したがって，歌声 F0 軌跡を単純に時間・周波数方向に離散化すると不自然なリズムや統計的に稀な半音階進行を含む音符系列がしばしば推定されてしまう．

この問題を解決するため，音符系列の生成過程を表現する楽譜モデルと歌声 F0 軌跡の生成過程を表現する F0 モデルの統合モデルに基づく調とリズムを考慮した統計的音符推定手法を提案する (図 1)．楽譜モデルでは，各小節ごとに調がマルコフ過程により決定され，音符の音高は調と直前の音符の音高に依存して決まる．各調のもとでは，調の音階構成音に含まれる音名を持つ音高が出やすくなるよう制御される．ビートグリッド上に定義される各音符の開始位置は，直前の音符の開始位置に依存することで，リズム構造を形成する．F0 モデルでは，楽譜モデルによって生成された楽譜に対応する階段関数状の F0 軌跡に時間・周波数方向の逸脱が付与される．統合モデルは階層隠れセミマルコフモデル (HHSMM: hierarchical hidden semi-Markov model) として定式化される．提案法は歌声 F0

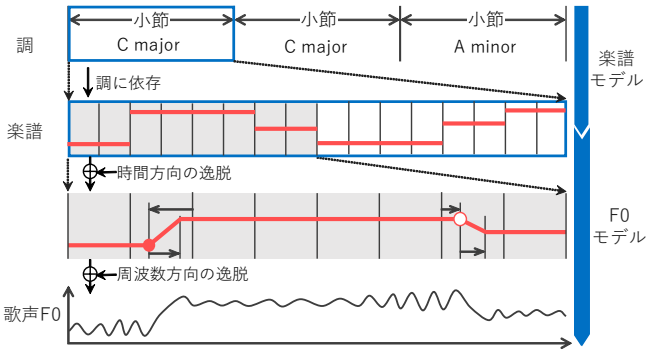


図 1: 楽譜モデルと F0 モデルの階層隠れセミマルコフモデルに基づく歌声 F0 軌跡の生成過程．

軌跡とビート時刻を入力として受け取り，マルコフ連鎖モンテカルロ法を用いることで，HHSMM の潜在変数として表現される調，音符，F0 の逸脱を同時に推定する．歌声 F0 軌跡の時間・周波数方向の離散化において調やリズムが自己組織化の制約として機能することが本手法の重要な特徴である．

2. 関 連 研 究

本章では歌声解析に関する研究を紹介する．

2.1 音楽音響信号に対する歌声 F0 推定

音楽音響信号に対する歌声 F0 軌跡の推定は活発に研究されており [1–5]，これらの出力結果は提案法の入力として用いられる．最も基本的な方法の 1 つとして，各 F0 候補のそれぞれについて高調波成分の和を計算する Subharmonic Summation (SHS) [1] がある．池宮ら [2] は SHS に基づく歌声 F0 推定とロバスト主成分分析 (RPCA: robust principal component analysis) に基づく歌声分離の性能を，これら 2 つのタスクの相互依存性を利用することで改善した．Salamon ら [12] は特徴関数を計算することで歌声 F0 軌跡の候補を推定し，各軌跡の特徴から主旋律を形成しない軌跡を再帰的に消去する手法を提案した．Durrieu ら [3] は歌声と伴奏をそれぞれソース・フィルターモデルと非負値行列因子分解 (NMF: non-negative matrix factorization) に基づくモデルで表現することにより，主旋律の分離を行った．Mauch ら [5] は YIN [4] を確率的な手法に修正することで，システムが複数の F0 候補を出力し，その中から各フレームごとに 1 つの F0 を HMM を用いて選択するようにした．

2.2 歌声に対する音符推定

歌声に対する音符推定も盛んに研究が行われている [11, 13–19]．素朴な手法として，一定の区間ごとに歌声 F0 の多数決をとって音符の音高を決定する手法がある [11]．Paiva ら [13] は多重音検出，複数の F0 軌跡の構築，それら軌跡の分割，不要な音符の消去，主旋律を形成する音符の抽出の 5 つの処理を順番に行う手法を提案した．Raphael [14] は音符の個数を与えて，音高，リズム，テンポを推定する HMM に基づく手法を提案した．我々の提案法で用いられているリズムや歌声の発音時刻の逸脱に関するモデルは [14] で用いられたものと同様である．Laaksonen ら [15] は入力として与えたコードの境界に注目することで音響データを調と音符に対応する区間に分割し，スコア関数に基づいて各区間ごとに音符を推定する手法を提案した．Ryynänen ら [16] は 1 つの音符内における種々の歌声変動 (例えば，ビブラートやポルタメント）を捉えるために階層 HMM に基づく手法を提案した．この手法のモデルでは，上層の HMM が音符の音高間の遷移を表し，下層の HMM が歌声変動の遷移を表す．Molina ら [17] は歌声 F0 軌跡における履歴現象に焦点を当てた．錦見ら [20] は時間・周波数方向の逸脱を考慮した歌声 F0 軌跡の生成過程を表現する HMM に基づく手法を提案した．Yang ら [18] は f_0-Δf_0 平面の生成過程を表現する階層 HMM に基づく手法を提案した．Mauch ら [19] は音高抽出を行う Tony というソフトウェアツールを開発した．このツール内では，PYIN [5] を用いて歌声 F0 推定を行い，Ryynänen's らの手法 [16] を基にした手法を用いて音符推定を行う．

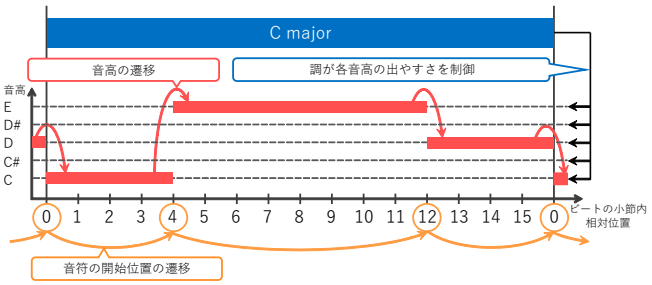


図 2: 楽譜モデル概要．

3. 提 案 法

本章では歌声 F0 軌跡から音符系列を推定する提案法について説明する．提案法は，調に依存する音符系列から時間・周波数方向の逸脱を伴って歌声 F0 軌跡が確率的に生成される過程を HHSMM に基づいて表現する．提案モデルの上層は小節に割り当てられた調に従って音符系列が確率的に生成される過程を表現する HMM である．下層は時間方向の逸脱と周波数方向の逸脱がそれぞれ潜在変数と出力確率として表現される HSMM である．

3.1 問 題 設 定

我々が取り組む問題を以下のように定める．

入力：歌声 F0 軌跡 $\boldsymbol{X}=\{x_t\}_{t=1}^{T}$ と 16 分音符単位のビート時刻 $\boldsymbol{Y}=\{(u_n, v_n)\}_{n=0}^{N}$,
出力：音符系列 $\boldsymbol{Z}=\{z_j=(p_j, l_j)\}_{j=0}^{J}$,

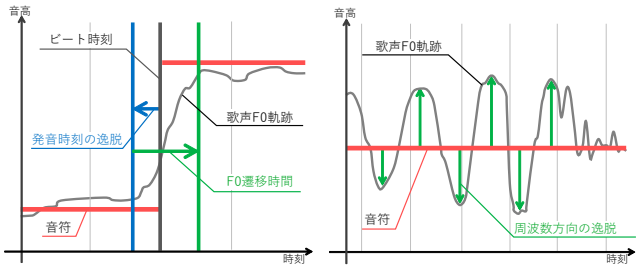
ここで，T は歌声 F0 軌跡のフレーム数，x_t は時刻 t における対数周波数，N は 16 分音符単位のビートの数である．$u_n \in \{1, \ldots, T{+}1\}$ は n 番目のビート時刻であり，楽曲の最初と最後は $u_0 = 1$ と $u_N = T{+}1$ としてそれぞれ表される．$v_n \in \{0, \ldots, 15\}$ は n 番目のビートが所属する小節内において，小節の先頭からそのビートまでの相対的な位置を表す．J は提案法によって推定される音符の個数であり，j 番目の音符 z_j は半音単位の音高 $p_j \in \{1, \ldots, K\}$ と 16 分音符単位の音価 $l_j \in \{1, \ldots, L\}$ の組として表現される．ここで，K は楽譜中に現れる音高の種類数であり，p_j は半音単位の音高に対応する対数周波数の集合 $\{\mu_1, \ldots, \mu_K\}$ のうちの 1 つを指し示す．初期音符 z_0 は便宜上導入された実際の楽譜には現れない音符である．

3.2 楽譜の確率的モデル化

本章では音符の音高が調とリズムに依存して生成される過程を表現する HMM に基づく楽譜モデルについて説明する．

3.2.1 調遷移のモデル化

調系列は $\boldsymbol{S} = \{s_m\}_{m=0}^{M}$ で表現され，M は楽曲中の小節数，s_m は m 番目の小節の調を表す．便宜上，初期音符 z_0 が所属する初期小節を導入し，その小節に割り当てられる調を s_0 とする．転調に対応できるようにするため曲全体で調を 1 つに固定



(a) 時間方向の逸脱　(b) 周波数方向の逸脱

図 3: 歌声 F0 軌跡に含まれる逸脱.

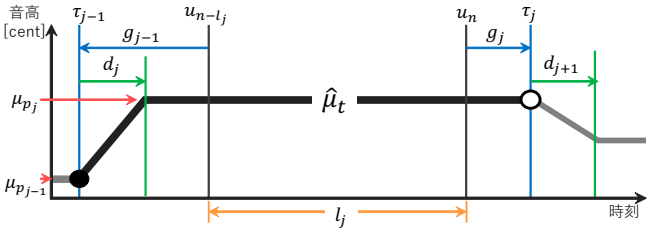


図 4: HSMM の潜在変数と出力確率の位置パラメータの関係. 黒い太線がコーシー分布の位置パラメータを表す.

せず，小節線で切り替わるようにする．各調 s_m は 24 通りの値 $\{\mathrm{C}, \mathrm{C\#}, \cdots, \mathrm{B}\} \times \{\text{major, minor}\}$ の中のいずれの値をとる．潜在変数 $\boldsymbol{S}$ は以下のようにマルコフ連鎖をなす．

$$p(s_0|\boldsymbol{\pi}) = \pi_{s_0} \tag{1}$$

$$p(s_m|s_{m-1}, \boldsymbol{\xi}_{s_{m-1}}) = \xi_{s_{m-1}s_m} \tag{2}$$

ここで，$\boldsymbol{\pi} \in \mathbb{R}^{24}_{\geqq 0}$ は初期確率，$\boldsymbol{\xi}_s \in \mathbb{R}^{24}_{\geqq 0}$ は遷移確率である．

3.2.2 音高遷移のモデル化

音高系列 $\boldsymbol{P}$ は以下のように調系列 $\boldsymbol{S}$ に依存したマルコフ連鎖によって生成される (図 2).

$$p(p_0|s_0, \boldsymbol{\phi}_{s_0}) = \phi_{s_0 p_0} \tag{3}$$

$$p(p_j|p_{j-1}, s_m, \boldsymbol{\psi}_{s_m p_{j-1}}) = \psi_{s_m p_{j-1} p_j} \tag{4}$$

ここで，$\boldsymbol{\phi}_s \in \mathbb{R}^K_{\geqq 0}$ は初期確率，$\boldsymbol{\psi}_{sp} \in \mathbb{R}^K_{\geqq 0}$ は遷移確率，m は音符 z_j が属する小節のインデックスである．さらに，$\phi_{s_0 p_0}$ と $\psi_{s_m p_{j-1} p_j}$ を以下のように定義する．

$$\phi_{s_0 p_0} = \frac{\hat{\phi}_{\hat{s}_0 \deg(p_0; s_0)}}{\sum_{p=1}^{K} \hat{\phi}_{\hat{s}_0 \deg(p; s_0)}} \tag{5}$$

$$\psi_{s_m p_{j-1} p_j} = \frac{\hat{\psi}_{\hat{s}_m \deg(p_{j-1}; s_m) \deg(p_j; s_m)}}{\sum_{p=1}^{K} \hat{\psi}_{\hat{s}_m \deg(p_{j-1}; s_m) \deg(p; s_m)}} \tag{6}$$

ここで，$\hat{s} \in \{\text{major,minor}\}$ は調 s の旋法，$\deg(p;s) \in \{0, \ldots, 11\}$ は調 s における音高 p の度数 (調 s の主音に対する p のピッチクラスの音程) である．$\hat{\phi}_*$ と $\hat{\psi}_*$ はそれぞれ旋法が与えられた下でのピッチクラスの初期確率と遷移確率である．

3.2.3 音符の開始位置遷移のモデル化

隣接する音符の開始位置間の遷移を考慮することで，音符系列 $\boldsymbol{Z}$ が妥当なリズムを持つようにする．j 番目の音符 z_j の開始位置を $r_{j-1} \in \{v_n\}_{n=1}^N$ とすると，音符の開始位置の遷移確率は以下のように与えられる．

$$p(r_j|r_{j-1}, \boldsymbol{\zeta}_{r_{j-1}}) = \zeta_{r_{j-1} r_j} \tag{7}$$

ここで，r_{j-1} と r_j との間の距離は音符 z_j の音価 l_j となる．曲の最初と最後に関しては $r_0 = v_0$ と $r_J = v_N$ とする．

3.3 歌声 F0 軌跡の確率的モデル化

本章では歌声 F0 軌跡の生成過程を表現する HSMM に基づく F0 モデルについて説明する．提案モデルでは，音符の音高と開始位置，時間方向の逸脱は潜在変数として表現され，周波数方向の逸脱は出力確率として表現する．

3.3.1 時間方向の逸脱のモデル化

歌声 F0 軌跡には以下のように 2 種類の時間方向の逸脱が含まれると仮定する (図 3a).

発音時刻の逸脱: 歌声の発音時刻と音符の開始位置との間のずれ.

F0 の遷移時間: ある音符の音高から次の音符の音高まで，歌声が遷移し切るのに要する時間.

音符系列 $\boldsymbol{Z}$ に付与される発音時刻の逸脱 $\boldsymbol{G} = \{g_j\}_{j=0}^J$ は離散潜在変数として表現される．音符の開始位置モデルと同様に音符 z_j の発音時刻の逸脱を g_{j-1} とする．各 g_j は $-G$ から G までの整数値を取り，以下のようにそれぞれ独立に生成されるとする．

$$p(g_j|\boldsymbol{\rho}) = \rho_{g_j} \tag{8}$$

ここで，$\boldsymbol{\rho} \in \mathbb{R}^{2G+1}_{\geqq 0}$ は発音時刻逸脱の確率の集合である．また，最初の音符の開始時刻と最後の音符の終了時刻には逸脱が無い，すなわち $g_0 = g_J = 0$ であるとする．

音符系列 $\boldsymbol{Z}$ に付与される F0 の遷移時間 $\boldsymbol{D} = \{d_j\}_{j=1}^J$ は離散潜在変数として表現され，各 d_j は 1 から D までの整数値をとる．音符 z_{j-1} と z_j の間における歌声 F0 軌跡の連続的な遷移は，幅が d_j フレームの斜め線によって表現される．各 d_j は以下のように独立に生成される．

$$p(d_j|\boldsymbol{\eta}) = \eta_{d_j} \tag{9}$$

ここで，$\boldsymbol{\eta} \in \mathbb{R}^D_{\geqq 0}$ は F0 の遷移時間の確率の集合である．

3.3.2 周波数方向の逸脱のモデル化

歌声 F0 軌跡 $\boldsymbol{X} = \{x_t\}_{t=1}^T$ は時間方向の逸脱が既に付与された音符系列に対して周波数方向の逸脱がさらに付与されて生成される (図 3b)．x_t は各フレームごとに独立に生成されるとし，j 番目の音符 z_j に関する出力確率は以下の通りである．

$$\begin{aligned} &p(x_{\tau_{j-1}:\tau_j-1}|p_{j-1}, p_j, l_j, g_{j-1}, g_j, d_j, \hat{\mu}_t, \lambda) \\ &= \prod_{t=\tau_{j-1}}^{\tau_j-1} \{\delta_{x_t,\text{voiced}} \mathrm{Cauchy}(x_t|\hat{\mu}_t, \lambda) + \delta_{x_t,\text{unvoiced}}\} \\ &= e_{p_{j-1} p_j l_j g_{j-1} g_j d_j} \end{aligned} \tag{10}$$

ここで，$x_{\tau':\tau-1}$ は $x_{\tau'}, \ldots, x_{\tau-1}$ を表し，λ は周波数方向の逸脱の大きさを表現する尺度パラメータ，δ はクロネッカーのデ

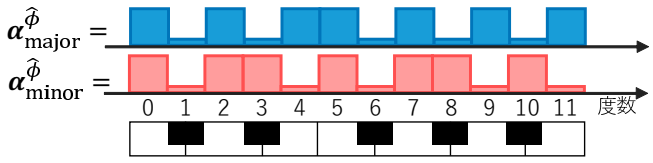


図 5: ハイパーパラメータ $\boldsymbol{a}^{\hat{\phi}}_{\hat{s}}$ の設定.

ルタ，$\hat{\mu}_t$ (図 4) は以下のように定義される位置パラメータである．

$$\hat{\mu}_t=\begin{cases}\frac{\mu_{p_j}-\mu_{p_{j-1}}}{d_j}(t-\tau_{j-1})+\mu_{p_{j-1}} & (\tau_{j-1}\leqq t<\tau_j+d_j)\\ \mu_{p_j} & (\tau_{j-1}+d_j\leqq t<\tau_j)\end{cases} \tag{11}$$

音符 z_{j+1} の開始位置が n 番目のビートに位置しているとき，$\tau_j = u_n + g_j$，$\tau_{j-1} = u_{n-l_j} + g_{j-1}$ である．

3.4 事前分布の導入

離散モデルパラメータ $\boldsymbol{\pi}$, $\boldsymbol{\xi}$, $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\psi}}$, $\boldsymbol{\zeta}$, $\boldsymbol{\rho}$,$\boldsymbol{\eta}$ に対して以下のようにディリクレ共役事前分布を置く．

$$\begin{aligned}
&\boldsymbol{\pi} \sim \text{Dirichlet}\left(\boldsymbol{a}^{\pi}\right) \quad \boldsymbol{\xi}_s \sim \text{Dirichlet}\left(\boldsymbol{a}^{\xi}_s\right)\\
&\hat{\boldsymbol{\phi}}_{\hat{s}} \sim \text{Dirichlet}\left(\boldsymbol{a}^{\hat{\phi}}_{\hat{s}}\right) \quad \hat{\boldsymbol{\psi}}_{\hat{s}\text{deg}(p;s)} \sim \text{Dirichlet}\left(\boldsymbol{a}^{\hat{\psi}}_{\hat{s}\text{deg}(p;s)}\right)\\
&\boldsymbol{\zeta}_r \sim \text{Dirichlet}\left(\boldsymbol{a}^{\zeta}_r\right)\\
&\boldsymbol{\rho} \sim \text{Dirichlet}\left(\boldsymbol{a}^{\rho}\right) \quad \boldsymbol{\eta} \sim \text{Dirichlet}\left(\boldsymbol{a}^{\eta}\right)
\end{aligned} \tag{12}$$

ここで，$\boldsymbol{a}^{\pi} \in \mathbb{R}^{26}_+$, $\boldsymbol{a}^{\xi}_s \in \mathbb{R}^{26}_+$, $\boldsymbol{a}^{\hat{\phi}}_{\hat{s}} \in \mathbb{R}^{12}_+$, $\boldsymbol{a}^{\hat{\psi}}_{\hat{s}\text{deg}(p;s)} \in \mathbb{R}^{12}_+$, $\boldsymbol{a}^{\zeta}_r \in \mathbb{R}^{16}_+$, $\boldsymbol{a}^{\rho} \in \mathbb{R}^{2G+1}_+$, $\boldsymbol{a}^{\eta} \in \mathbb{R}^{D}_+$ はハイパーパラメータである．ある調のもとでの各ピッチクラスの出やすさは，それらピッチクラスの初期確率と遷移確率に関する事前分布を用いて制御される．図 5 に示すように，ハイパーパラメータ $\boldsymbol{a}^{\hat{\phi}}_{\hat{s}}$ と $\boldsymbol{a}^{\hat{\psi}}_{\hat{s}\text{deg}(p;s)}$ はそれぞれダイアトニックスケールを表現するように設定される．コーシー分布は共役事前分布を持たないので，尺度パラメータ λ に対して以下のようにガンマ事前分布を置く．

$$\lambda \sim \text{Gamma}\left(a^{\lambda}_0, a^{\lambda}_1\right) \tag{13}$$

ここで，a^{λ}_0 と a^{λ}_1 はハイパーパラメータである．

3.5 ベイズ推定

我々の目的は歌声 F0 軌跡 $\boldsymbol{X}$ が与えられた下で事後分布 $p(\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Theta}|\boldsymbol{X})$ を計算することである．ここで，$\boldsymbol{Q} = \{\boldsymbol{P}, \boldsymbol{L}, \boldsymbol{G}, \boldsymbol{D}\}$ (潜在変数)，$\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \boldsymbol{\xi}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}}, \boldsymbol{\zeta}, \boldsymbol{\rho}, \boldsymbol{\eta}, \lambda\}$ (モデルパラメータ) である．この事後分布は解析的に計算することが困難であるため，我々はマルコフ連鎖モンテカルロ法 (MCMC: Markov chain Monte Carlo) を用いて $\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Theta}$ の値をサンプルする．潜在変数 $\boldsymbol{S}$ と $\boldsymbol{Q}$ のサンプルに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アルゴリズムを用いる．モデルパラメータ $\boldsymbol{\Theta}$ のうち λ 以外，すなわち共役事前分布を持つパラメータ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．パラメータ λ には共役事前分布が無いので，メトロポリス・ヘイスティングス (MH: Metropolis-Hastings) アルゴリズムを用いる．$\boldsymbol{S}$ と $\boldsymbol{Q}$ は音符系列 $\boldsymbol{Z}$ を共有し，相互に依存しているため，各変数は以下の手順で更新される．

（1） 多数決法により音符系列 $\boldsymbol{Z}$ を初期化する．
（2） $\boldsymbol{Z}$ に基づき調系列 $\boldsymbol{S}$ を更新する．
（3） $\boldsymbol{S}$ に基づき $\boldsymbol{Q}$ を更新する．
（4） モデルパラメータ $\boldsymbol{\Theta}$ を更新する．
（5） 2 に戻る．

3.5.1 潜在変数 $\boldsymbol{S}$ の推論

音符系列 $\boldsymbol{Z}$ が与えられた下で各 s_m は以下に示す確率に従いサンプルされる．

$$\beta^S_{s_m} = p(s_m|s_{m+1:M}, \boldsymbol{Z}) \tag{14}$$

ここで，$s_{m+1:M}$ は $s_{m+1}, \ldots, s_M$ を表す．式 (14) の計算と調 $\boldsymbol{S}$ のサンプルに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法を用いる．

フォワードフィルタリングでは確率 $\alpha^S_{s_m}$ が以下のように再帰的に計算される．

$$\alpha^S_{s_0} = p(p_0, s_0) = p(p_0|s_0)p(s_0) = \phi_{s_0 p_0}\pi_{s_0} \tag{15}$$

$$\begin{aligned}\alpha^S_{s_m} &= p(p_{0:j_{m+1}-1}, s_m)\\ &= \prod_{j=j_m}^{j_{m+1}-1}\psi_{s_m p_{j-1} p_j}\sum_{s_{m-1}}\xi_{s_{m-1}s_m}\alpha^S_{s_{m-1}}\end{aligned} \tag{16}$$

ここで，j_m は m 番目の小節に属する最初の音符のインデックスであり，既知の音価 $\boldsymbol{L}$ から計算できる．

バックワードサンプリングではフォワードフィルタリングで計算された値を用いて式 (14) を計算し，以下のように調が再帰的にサンプルされる．

$$\beta^S_{s_M} = p(s_M|\boldsymbol{Z}) \propto \alpha^S_{s_M} \tag{17}$$

$$\beta^S_{s_m} = p(s_m|s_{m+1:M}, \boldsymbol{Z}) \propto \alpha^S_{s_m}\xi_{s_m s_{m+1}} \tag{18}$$

3.5.2 潜在変数 $\boldsymbol{Q}$ の推論

潜在変数 $\boldsymbol{Q}$ は $\boldsymbol{S}$ と同様の方法で推論される．フォワードフィルタリングでは確率 $\alpha^Q_{p_n l_n, g_n d_n}$ が以下のように再帰的に計算される．

$$\alpha^Q_{p_0 l_0 g_0 d_0} = p(p_0|\boldsymbol{S}) = \phi_{y_0 p_0} \tag{19}$$

$$\begin{aligned}\alpha^Q_{p_n l_n g_n d_n} &= p(x_{1:\tau_n-1}, p_n, l_n, g_n, d_n|\boldsymbol{S})\\ &= \begin{cases}0 & (l_n > n)\\ \rho_{g_n}\eta_{d_n}\zeta_{r_0 r_n} & \\ \quad\cdot\sum_{p_0}\psi_{s_1 p_0 p_n}e_{p_0 p_n l_n 0 g_n d_n}\alpha^Q_{p_0 l_0 g_0 d_0} & (l_n = n)\\ \sum_{p_{n'}, g_{n'}}\sum_{l_{n'}}^{\min(n', L)}\sum_{d_{n'}}\rho_{g_n}\eta_{d_n}\zeta_{r_{n'} r_n}\psi_{s_{m(n')} p_{n'} p_n} & \\ \quad\cdot e_{p_{n'} p_n l_n g_{n'} g_n d_n}\alpha^Q_{p_{n'} l_{n'} g_{n'} d_{n'}} & (l_n < n)\end{cases}\end{aligned} \tag{20}$$

ここで，$\tau_n = u_n + g_n$, $n' = n - l_n$ であり，$m(n')$ は n' 番目のビートが属する小節のインデックスである．p_n, l_n, g_n, d_n は終了位置が n 番目のビート u_n に位置する音符に対応するフォワードメッセージの変数である．これらの変数は j を添字とする変数 p_j, l_j, g_j, d_j とは異なる．音符 $z_n = (p_n, l_n)$ の開始位置と終了位置はそれぞれ $(n-l_n)$ 番目のビートと n 番目のビートに位置し，式 (20) の再帰計算に現れる確率 $p(l_n)$ は

$p(r_n|r_{n-l_n})$ に置き換えられる．

バックワードサンプリングではフォワードフィルタリングで計算された値を用いて潜在変数の事後分布を計算し，以下のように音符と時間方向の逸脱が再帰的にサンプルされる．

$$\begin{aligned}
&\beta_{p_N l_N g_N d_N} = p(p_N, l_N, g_N, d_N|\boldsymbol{X}, \boldsymbol{S}) \propto \alpha^Q_{p_N l_N g_N d_N} \\
&\beta_{p_{n'} l_{n'} g_{n'} d_{n'}} \\
&= p(p_{n'}, l_{n'}, g_{n'}, d_{n'}|p_{n:N}, l_{n:N}, g_{n:N}, d_{n:N}, \boldsymbol{X}) \\
&\propto \begin{cases} 0 & (l_n > n) \\ e_{p_{n'} p_n l_n g_{n'} g_n d_n} \psi_{s_{m(n')} p_{n'} p_n} & \\ \quad \cdot \zeta_{r_{n'} r_n} \rho_{g_n} \eta_{d_n} \alpha^Q_{p_{n'} l_{n'} g_{n'} d_{n'}} & (l_n \leqq n) \end{cases}
\end{aligned} \tag{21}$$

3.5.3 モデルパラメータ $\boldsymbol{\Theta}$ の学習

共役事前分布を持つモデルパラメータの事後分布はバックワードダンプリングで得られたサンプル $\boldsymbol{S}$，$\boldsymbol{Q}$ を用いて計算される．そして，これらのパラメータは計算された事後分布に従い以下のようにサンプルされる．

$$\boldsymbol{\pi} \sim \mathrm{Dirichlet}\left(\boldsymbol{a}^{\pi} + \boldsymbol{b}^{\pi}\right) \quad \boldsymbol{\xi}_s \sim \mathrm{Dirichlet}\left(\boldsymbol{a}^{\xi}_s + \boldsymbol{b}^{\xi}_s\right) \tag{22}$$

$$\hat{\boldsymbol{\phi}}_{\hat{s}} \sim \mathrm{Dirichlet}\left(\boldsymbol{a}^{\hat{\phi}}_{\hat{s}} + \boldsymbol{b}^{\hat{\phi}}_{\hat{s}}\right) \tag{23}$$

$$\hat{\boldsymbol{\psi}}_{\hat{s}\mathrm{deg}(p;s)} \sim \mathrm{Dirichlet}\left(\boldsymbol{a}^{\hat{\psi}}_{\hat{s}\mathrm{deg}(p;s)} + \boldsymbol{b}^{\hat{\psi}}_{\hat{s}\mathrm{deg}(p;s)}\right) \tag{24}$$

$$\boldsymbol{\zeta}_r \sim \mathrm{Dirichlet}\left(\boldsymbol{a}^{\zeta}_r + \boldsymbol{b}^{\zeta}_r\right) \tag{25}$$

$$\boldsymbol{\rho} \sim \mathrm{Dirichlet}\left(\boldsymbol{a}^{\rho} + \boldsymbol{b}^{\rho}\right) \quad \boldsymbol{\eta} \sim \mathrm{Dirichlet}\left(\boldsymbol{a}^{\eta} + \boldsymbol{b}^{\eta}\right) \tag{26}$$

ここで，$\boldsymbol{b}^{\pi} \in \mathbb{R}^{26}_{\geqq 0}$ は第 s_0 成分だけが 1 の単位ベクトルである．$\boldsymbol{b}^{\xi}_s \in \mathbb{R}^{26}_{\geqq 0}$ はサンプル $\boldsymbol{Y}$ における調 s から s' への遷移の回数を第 s' 成分が表すベクトルである．$\boldsymbol{b}^{\rho} \in \mathbb{R}^{2G+1}_{\geqq 0}$ はサンプル $\boldsymbol{Q}$ における歌声の発音時刻の逸脱が g である回数を第 g 成分が表すベクトルであり，$\boldsymbol{b}^{\eta} \in \mathbb{R}^{D}_{\geqq 0}$ はサンプル $\boldsymbol{Q}$ における F0 の遷移時間が d である回数を第 d 成分が表すベクトルである．$\boldsymbol{b}^{\zeta}_r \in \mathbb{R}^{16}_{\geqq 0}$ は，バックワードサンプリングで得られた音価 $\boldsymbol{L}$ から計算される $\boldsymbol{R} = \{r_j\}^J_{j=0}$ において，音符の開始位置 r から r' への遷移の回数を第 r' 成分が表すベクトルである．ベクトル $\boldsymbol{b}^{\hat{\phi}}_{\hat{s}} \in \mathbb{R}^{12}_{>0}$ は，初期小節の調と初期音符の音高がそれぞれ $s_0 = s$，$p_0 = p$ である時，要素 $b^{\hat{\phi}}_{\hat{s}\mathrm{deg}(p;s)}$ の値が 1 でそれ以外の要素の値が 0 であるようなベクトルである．ベクトル $\boldsymbol{b}^{\hat{\psi}}_{\hat{s}\mathrm{deg}(p;s)} \in \mathbb{R}^{12}_{\geqq 0}$ は，サンプルされた潜在変数において調 s の下で音高 p から p' への遷移の回数を要素 $b^{\hat{\psi}}_{\hat{s}\mathrm{deg}(p;s)\mathrm{deg}(p';s)}$ が表すベクトルである．

尺度パラメータ λ については MH アルゴリズムを適用するため，以下のように提案分布を定める．

$$q(\lambda^*|\lambda) = \mathrm{Gamma}(\gamma\lambda, \gamma) \tag{27}$$

ここで，λ^* は次の λ の値の候補を表す変数，λ は現在の尺度パラメータの値，γ はハイパーパラメータである．λ の値の更新は以下の確率に従い行われる．

$$A(\lambda^*, \lambda) = \min\left\{\frac{\mathcal{L}(\lambda^*) q(\lambda|\lambda^*)}{\mathcal{L}(\lambda) q(\lambda^*|\lambda)}\right\} \tag{28}$$

ここで，$L(\lambda)$ は以下のように与えられる．

$$\mathcal{L}(\lambda) = \mathrm{Gamma}\left(\lambda|a^{\lambda}_0, a^{\lambda}_1\right) \prod_{j=1}^{J} e_{p_{j-1} p_j l_j g_{j-1} g_j d_j} \tag{29}$$

$\{p_j, l_j, g_j, d_j\}^J_{j=0}$ はバックワードサンプリングでサンプルされた値である．一様分布 $\mathcal{U}(0,1)$ からサンプルされた乱数よりも採択率 $A(\lambda^*, \lambda)$ の値が大きい場合に λ の値は λ^* に更新される．

3.6 ビタビ復号

潜在変数系列 $\boldsymbol{S}$，$\boldsymbol{Q}$ は学習時に同時分布 $p(\boldsymbol{X}, \boldsymbol{Q}, \boldsymbol{S}, \boldsymbol{\Theta}|\boldsymbol{\Phi})$ を最大化したモデルパラメータの値を用いるビタビアルゴリズムによって推定される．潜在変数の推論と同様に，多数決法により音符系列 $\boldsymbol{Z}$ を初期化したのち，$\boldsymbol{Z}$ に基づき $\boldsymbol{S}$ を推定し，推定した $\boldsymbol{S}$ に基づき $\boldsymbol{Q}$ を推定する．

$\boldsymbol{S}$ に関するビタビ復号では，以下のように ω^S_s が再帰的に計算される．

$$\omega^S_{s_0} = \ln \phi_{s_0 k_0} + \ln \pi_{s_0} \tag{30}$$

$$\omega^S_{s_m} = \sum_{j=j_m}^{j_{m+1}-1} \ln \psi_{s_m p_{j-1} p_j} + \max_{s_{m-1}} \left\{\ln \xi_{s_{m-1} s_m} + \omega^S_{s_{m-1}}\right\} \tag{31}$$

ω^S_s の再帰計算では，$\omega^S_{s_m}$ の値を最大化する 1 つ前の状態 s_{m-1} が $c^S_{s_m}$ として記録され，調系列 $\boldsymbol{S}$ は以下のように再帰的に推定される．

$$s_M = \arg\max_{s_M} \alpha^S_{s_M} \quad s_{m-1} = c^S_{s_m} \tag{32}$$

$\boldsymbol{Q}$ に関するビタビ復号では，ω^Q_{plgd} の値が以下のように再帰的に計算される．

$$\omega^Q_{p_0 l_0 g_0 d_0} = \mathrm{w}^{\phi} \ln \phi_{s_0 p_0} \tag{33}$$

$$\begin{aligned}
&\omega^Q_{p_n l_n g_n d_n} \\
&= \begin{cases} -\inf & (l_n > n) \\ \mathrm{w}^{\rho} \ln \rho_{g_n} + \mathrm{w}^{\eta} \ln \eta_{d_n} + \mathrm{w}^{\zeta} \ln \zeta_{r_n r_0} & \\ \quad + \max_{p_0}\Big\{\mathrm{w}^{\psi} \ln \psi_{s_1 p_0 p_n} & \\ \quad + \mathrm{w}^{e} \ln e_{p_0 p_n l_n 0 g_n d_n} + \omega^Q_{p_0 l_0 g_0 d_0}\Big\} & (l_n = n) \\ \mathrm{w}^{\rho} \ln \rho_{g_n} + \mathrm{w}^{\eta} \ln \eta_{d_n} + \mathrm{w}^{\zeta} \ln \zeta_{r_n r_{n'}} & \\ \quad + \max_{(p_{n'}, l_{n'}, g_{n'}, d_{n'})}\Big\{\mathrm{w}^{\psi} \ln \psi_{s_{m(n')} p_{n'} p_n} & \\ \quad + \mathrm{w}^{e} \ln e_{p_{n'} p_n l_n g_{n'} g_n d_n} + \omega^Q_{p_{n'} l_{n'} g_{n'} d_{n'}}\Big\} & (l_n < n) \end{cases}
\end{aligned} \tag{34}$$

ここで，w^{ϕ}, w^{ψ}, w^{ρ}, w^{η}, w^{ζ}, w^{e} は各確率間のバランスを制御する重みパラメータである．ω^Q_{plgd} の再帰計算では，$\omega^Q_{p_n l_n g_n d_n}$ の値を最大化する 1 つ前の状態 $p_{n'}$, $l_{n'}$, $g_{n'}$, $d_{n'}$ が $c^Q_{p_n l_n g_n d_n}$ として記録され，$\boldsymbol{Q}$ は以下のように再帰的に推定される．

$$(p_N, l_N, g_N, d_N) = \arg\max_{p_N, l_N, g_N, d_N} \alpha^Q_{p_N l_N g_N d_N} \tag{35}$$

$$(p_{n'}, l_{n'}, g_{n'}, d_{n'}) = c^Q_{p_n l_n g_n d_n} \tag{36}$$

4. 評価実験

歌声 F0 軌跡からの音符推定について，提案法の精度を評価

表 1: ビート単位と音符単位の一致率 [%] および標準誤差

モデル	入力 F0	ビート単位	音符単位
提案法	正解	**72.4** ± 1.7	28.1 ± 2.1
	推定	68.7 ± 1.3	30.7 ± 1.8
リズムのみ考慮	正解	71.5 ± 1.6	26.3 ± 2.1
	推定	67.7 ± 1.3	29.1 ± 1.8
調のみ考慮	正解	67.8 ± 1.6	10.6 ± 1.2
	推定	65.6 ± 1.2	13.8 ± 1.1
調・リズムを考慮しない	正解	67.2 ± 1.5	9.8 ± 1.2
	推定	64.6 ± 1.2	12.9 ± 1.1
多数決法	正解	54.1 ± 1.5	20.1 ± 1.4
	推定	61.0 ± 1.4	22.0 ± 1.5
HMM [20]	推定	68.0 ± 1.2	14.8 ± 1.3

するために比較実験を行った．

4.1 実験条件

RWC 研究用音楽データベース [21] のポピュラー音楽 100 曲のうち，提案法が扱えない 32 分音符，3 連符，ハモリパートなどを含む曲を除いた 63 曲を用いた．入力の歌声 F0 軌跡 $\boldsymbol{X}$ はアノテーションデータ [22] と [2] で提案されている手法よって推定されたものを用いた．アノテーションデータには無声区間が含まれるが推定データには含まれない．ビート時刻とビートの小節内における相対位置 $\boldsymbol{Y}$ はアノテーションデータから得た．

ベイズ推定とビタビ復号は各曲独立に行われる．音符の開始位置の遷移確率はロック音楽のコーパス [23] から事前に学習した．ハイパーパラメータは $\boldsymbol{a}^{\pi}$=$\mathbf{1}$, $\boldsymbol{a}^{\xi}_{\hat{s}}$=$\mathbb{1}$, $\boldsymbol{a}^{\zeta}_{r}$=$\mathbf{1}$, $\boldsymbol{a}^{\rho} = \boldsymbol{a}^{\eta} = a^{\lambda}_{0}$ $= a^{\lambda}_{1} = \gamma$ =1, とした．ここで，$\mathbb{1}$ と $\mathbf{1}$ はそれぞれすべての要素が 1 の行列とベクトルである．$\boldsymbol{a}^{\hat{\phi}}_{\hat{s}}$ と $\boldsymbol{a}^{\hat{\psi}}_{\hat{s}\mathrm{deg}(p;s)}$ は，調 $\hat{s}$ のダイアトニックスケール内のピッチクラスに対応する要素を 10，それ以外の要素を 1 に設定した．ビタビアルゴリズムの重みパラメータは経験的に $\mathrm{w}^{\phi} = \mathrm{w}^{\psi} = 29.4$, $\mathrm{w}^{\rho} = 2.4$, $\mathrm{w}^{\eta} = 2.9$, $\mathrm{w}^{\zeta} = 48.5$, $\mathrm{w}^{e} = 3.8$ とした．音楽的に一貫性のある音符系列を得るために，F0 モデル関する重みパラメータよりも楽譜モデルに関する重みパラメータの値を大きくした．

比較ために，多数決法と HMM に基づく従来法 [20] についても実験した．従来法 [20] については，無声区間を含む歌声 F0 軌跡を入力として扱えないため，無音区間を含まない推定データについてのみ実験した．言語モデルの有効性を評価するために，提案法に関しても，1) 調とリズムのどちらも考慮しない手法，2) 調のみを考慮した手法，3) リズムのみを考慮した手法，4) 調とリズムの両方を考慮した手法，の 4 通りの実験を行った．提案法における学習を高速化するため，音符の音高の探索範囲を多数決法によって推定された音高の周囲に限定した．

各手法の性能を評価するため，正解の音符系列と推定された音符系列を比較して，ビート単位の一致率と音符単位の一致率を計算した．ビート単位の一致率は正解の楽譜内の音符が存在するビート区間の個数に対して，正しく音高が推定されたビート区間の個数の割合とする．音符単位の一致率は正解楽譜内の音符の個数に対して，音高，開始位置，終了位置の 3 つ全てが正しく推定された音符の個数の割合とする．正解楽譜の中の隣接する音符が，同じ音高であるかタイで結ばれている場合，それらの音符をまとめて 1 つの音符とみなした．従来法 [20] では 16 分音符ごとに音高を出力するので，連続する同じ音高の並びを 1 つの音符とみなした．

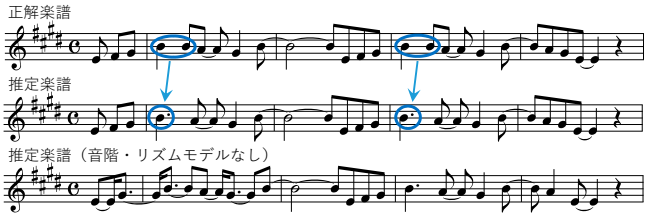


図 6: 提案法と調とリズムのどちらも考慮しない手法を用いて正解の歌声 F0 軌跡から推定された楽譜の例．

4.2 実験結果

実験結果を表 1 に示した．提案法は両評価尺度において，多数決法や従来法よりも音符推定精度が上回っていた．提案法に関する 4 通りの実験から得られたビート単位の一致率を比較すると，楽譜モデルによって音符推定の性能が向上することを確かめた．特に，調の遷移確率 (調の制約) よりも音符の開始時刻の遷移確率 (リズムの制約) の方がより有効であることが分かった．ビート単位の一致率では，提案法 (68.7%) と従来法 (68.0%) で大きな差は見られなかったが，音符単位の一致率では，提案法が (30.7%) が従来法 (14.8%) を大きく上回った．

推定された楽譜の例を図 6 に示す．一部の音符が結合されたことを除いて，提案法により推定された楽譜がほぼ正確であることがわかる．隣接する同じ音高の音符を正しく推定するには歌声 F0 軌跡だけでは限界があり，元の歌声や音楽音響信号を参照する必要がある．一方で，楽譜モデルを考慮せずに推定された楽譜には多くの推定誤りが含まれていた．この結果からも，音符推定において楽譜モデルを音楽的制約として用いることの有効性が示せた．

5. おわりに

本稿では歌声 F0 軌跡から音符系列を推定する統計的手法を示した．提案法は調から楽譜が生成される過程を表す楽譜モデルと楽譜から時間・周波数方向の逸脱を伴って歌声 F0 軌跡が生成される過程を表す F0 モデルを統合した階層隠れセミマルコフモデル (HHSMM) に基づいており，音楽的に一貫性のある音符系列を出力できることを確かめた．

本研究の今後の方向として最も興味深いのは，音楽音響信号に対する歌声 F0 推定において歌声 F0 軌跡の音楽的に有意な事前分布として提案したモデルを用いることである．本稿で提案した楽譜から F0 軌跡を生成するモデルを「言語」モデルとし，歌声 F0 軌跡からスペクトログラムを出力する音響モデルと階層ベイズの枠組みで統合する予定である．これにより音楽音響信号から歌声 F0 軌跡と楽譜を同時に学習することが可能になる．また，提案法では事前に推定した歌声 F0 軌跡とビート時

刻を入力として与えているが，入力の推定精度が音符推定精度に影響を与える問題を克服するためにも歌声 F0 軌跡とビート時刻の同時推定法について検討すべきである．

謝辞：本研究の一部は，JSPS 科研費 26700020, 16H01744, 16J05486 および JST ACCEL No. JPMJAC1602 の支援を受けた．

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 2015), pp.574–578, 2015.

[3] 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.

[4] 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.

[5] 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 2014), pp.659–663, 2014.

[6] 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.

[7] 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 2012), pp.57–60, 2012.

[8] Y.E. Kim and B. Whitman, “Singer identification in popular music recordings using voice coding features,” 3rd International Conference on Music Information Retrieval (ISMIR 2002), vol.13, p.17, 2002.

[9] 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.

[10] 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, pp.1417–1420, 2008.

[11] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, “Songle: A web service for active music listening improved by user contributions.,” Proc. of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), pp.311–316, 2011.

[12] 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.

[13] R.P. Paiva, T. Mendes, and A. Cardoso, “On the detection of melody notes in polyphonic audio.,” 6th International Conference on Music Information Retrieval (ISMIR 2005), pp.175–182, 2005.

[14] C. Raphael, “A graphical model for recognizing sung melodies.,” 6th International Conference on Music Information Retrieval (ISMIR 2005), pp.658–663, 2005.

[15] A. Laaksonen, “Automatic melody transcription based on chord transcription.,” Proc. of the 15th International Society for Music Information Retrieval (ISMIR 2014), pp.119–124, 2014.

[16] 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.

[17] E. Molina, L.J. Tardón, A.M. Barbancho, and I. Barbancho, “Sipth: Singing transcription based on hysteresis defined on the pitch-time curve,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol.23, no.2, pp.252–263, 2015.

[18] L. Yang, A. Maezawa, J.B.L. Smith, and E. Chew, “Probabilistic transcription of sung melody using a pitch dynamic model,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), pp.301–305, 2017.

[19] 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,” Proc. of the 1st International Conference on Technologies for Music Notation and Representation (TENOR 2015), pp.23–30, 2015.

[20] R. Nishikimi, E. Nakamura, K. Itoyama, and K. Yoshii, “Musical note estimation for f0 trajectories of singing voices based on a bayesian semi-beat-synchronous hmm,” Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016), pp.461–467, 2016.

[21] 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 2002), pp.287–288, 2002.

[22] M. Goto, “Aist annotation for the RWC music database.,” The 7th International Conference on Music Information Retrieval (ISMIR 2006), pp.359–360, 2006.

[23] T. De Clercq and D. Temperley, “A corpus analysis of rock harmony,” Popular Music, vol.30, no.01, pp.47–70, 2011.