Vol.2013-MUS-100 No.20
2013/9/1

伴奏付き歌唱に含まれる歌い方要素の個別抽出

池宮 由楽[1,a)]　糸山 克寿[1,b)]　奥乃 博[1,c)]

概要：本稿では，伴奏付き歌唱に含まれるビブラートやこぶしといった歌い方要素を個別に抽出する手法について述べる．歌い方要素は歌唱者の個人性を強く反映し，それらを個別に検出しパラメータ化することで，CGM や MIR への多様な応用が可能となる．本手法では，ユーザが簡易に取得できる歌唱の音高列を事前知識として用いる．音高列から探索範囲を制限したビタビ探索によって高精度に F0 を推定する．各要素は歌唱者の意図による F0 の特徴的な変動として現れ，それらを個別に検出し，設計したモデルに従ってパラメータとして抽出する．評価実験により，市販楽曲からプロ歌手の歌い方要素を個別に抽出できることを確認した．

1. はじめに

歌手はそれぞれ独自の歌い方（アレンジの癖）を持っており，それが個性となり歌手自身の魅力となっている．本研究の目的はこの歌手固有の歌い方をライブラリ化することであり，それによって近年盛んな CGM (Consumer Generated Media) や MIR (Music Information Retrieval) への活用を担う．例えば，ライブラリを用いて，特定歌手の歌い方を VOCALOID などを用いた合成歌唱へ転写したり，自分の好きな歌手と似た歌い方の歌手や演奏が検索できるようになる．

歌い方というのは抽象的なものであるため，転写や検索などを行うためには，何らかの枠組みに落としこみ，パラメータとして保存する必要がある．従来の歌い方を扱う技術には，ユーザ歌唱を VOCALOID に転写する VocaListener [1]，２つの歌唱の声質と歌い回しをモーフィングし合成を行う v.morish [2]，HMM 音声合成技術 Sinsy [3] がある．しかし，VocaListener，v.morish では音量や音高変化全体を歌い方として捉えるのみで，その特徴を分析しているわけではない．Sinsy では，HMM により学習する特徴ベクトルに歌い方が含まれるが，学習に多量の歌声と対応した楽譜が必要であり，様々な歌手の歌い方のライブラリ化は容易ではない．大石ら [4] は F0 軌跡から確率モデルを用いて歌い方に関する成分を推定したが，その成分の中での分析は行なっていない．

本稿では，歌唱者の意図によって付加される歌唱表現であるビブラート・グリッサンド・こぶし（小節）を「歌い方要素」として，それぞれ個別に抽出する手法について述べる．本稿で対象とする歌い方要素は，全て歌唱の F0（基本周波数）の変動として現れるものに限定し，抽出された歌い方要素は，本稿で設計された表現に落としこみパラメータ化される．また，ユーザのニーズとして市販楽曲に含まれるプロ歌手の分析がほとんどであると想定されるため，特に伴奏付き歌唱を対象として分析を行う．

伴奏付き歌唱からの歌い方要素抽出には次の２つの課題がある．

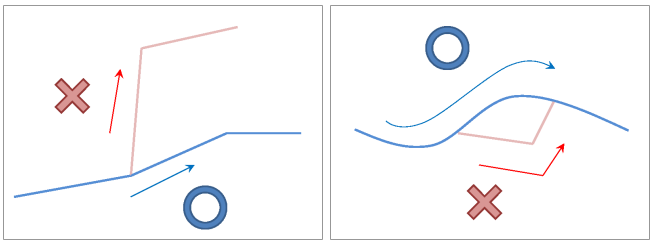
(1) 伴奏付き歌唱の F0 推定
(2) 歌唱 F0 からの歌い方要素の抽出・パラメータ化

伴奏付き歌唱の自動 F0 推定は困難な問題である．この問題を解決するため多くの手法 [5–7] が提案されているが，本稿では特に，前処理として歌声分離を行うことで歌唱 F0 推定に特化し，また，入力音高列により F0 を探索する周波数範囲を制限することで，推定の高精度化を実現する．F0 系列は，歌唱 F0 の特徴を確率的に取り入れ設計したマルコフモデルに基づきビタビ探索によって計算される．

歌唱 F0 から，各歌い方要素は時間的に重ならないという仮定のもと，設計したルールに従って順に抽出する．他の歌い方要素への誤抽出を抑制するため，なるべく一意に決定される要素から順に抽出していく．本稿では，ビブラート・グリッサンド・こぶしの順である．

本手法には各歌い方要素を個別に扱えることと，ユーザが市販楽曲を分析できるメリットがある．様々な歌手の異なる歌い方要素を組み合わせた歌声合成を行ったり，システムを CGM 化し様々な全国のユーザが協力して歌い方ライブ

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
[a)] ikemiya@kuis.kyoto-u.ac.jp
[b)] itoyama@kuis.kyoto-u.ac.jp
[c)] okuno@kuis.kyoto-u.ac.jp



(a) $\Delta F0$ (b) $\Delta\Delta F0$

図 1 歌声 F0 の時間的制約

ラリを作る仕組みを構築するなどの応用が考えられる．実験では，実際に市販楽曲から歌い方要素を抽出できることを確認する．

2. 問題設定

本稿で扱う問題をまとめると以下のようになる．

入力：	伴奏付き歌唱 / 歌唱音高列
出力：	歌唱に含まれる歌い方要素（パラメータ）
前提：	無声区間既知
処理単位：	音高を 10 – 20 程度含む素片 （A メロ・サビなど）

ここで歌唱音高列とは，{ ド・ミ・レ・ …} というように，伴奏付き歌唱に対応する楽譜の音高の並びを表した列である．歌唱音高列は，音符の長さなどが既知である必要のある楽譜自体に比べて，音を聴くだけで素人でも容易に取得できる情報である．また楽曲全体を分析するのではなく，ユーザが A メロ・サビといった分析に適した部分を選択することを想定している．

3. 伴奏音歌唱の F0 推定

伴奏付き歌唱の F0 推定は，以下の 3 つの処理からなる．

(1) 歌声分離．
(2) 音高列による F0 探索範囲制限．
(3) 歌声 F0 の時間的制約を考慮した周波数系列探索．

まず前処理として歌声分離を行うことで伴奏音の影響を抑制し，入力音高列によって F0 の探索範囲を制限することで推定の高精度化を実現する．本稿では歌声分離手法として REPET-SIM [8] を用い，探索範囲は $\min$(音高列)$-400[cent]$ から $\max$(音高列)$+400[cent]$ とする．REPET-SIM はブラインド歌声分離として最新であるとともに，伴奏抑制の影響で歌声が消える現象がほとんど起こらない．本稿で扱う問題は，全ての歌声区間が必要であるため，この手法が適切であると考えられる．続く節で，歌声 F0 の時間的制約を考慮した系列探索のモデル化・計算について述べる．

3.1 F0 推定の定式化

F0 推定は時間周波数領域で考えた場合，最も F0 らしい周波数の時系列を探索する問題と考えることができる．ここで，時間フレーム t で周波数 f がどれくらいの確率で F0 かを表した F0 尤度 $P_L(f_t)$ を導入する．最も単純には各時間フレームで F0 尤度が最大のものを F0 として推定すればよいが，これでは他楽器の F0 や倍ピッチの推定誤りが多く起こってしまう．

そこで，歌声 F0 系列の持つ特徴を時間的な制約として確率的に取り入れる．具体的には以下の 2 つに制約を与える．

- $\Delta F0$
- $\Delta\Delta F0$

ここで，$\Delta F0$ は歌声 F0 の急激に変化しないという特徴に相当し，$\Delta\Delta F0$ は歌声 F0 の滑らかに変化するという特徴に相当する (Figure 1)．

$\Delta F0$ と $\Delta\Delta F0$ の確率関数を，それぞれ $P_{\Delta F0}(f)$, $P_{\Delta\Delta F0}(f)$ とすると，F0 推定は次式を最大化する周波数系列 $\hat{F}$ を求める問題となる．

$$\hat{F} = \underset{F:=f_1,\ldots,f_T}{\arg\max} \left\{ \sum_{t=1}^{T} \log P_L(f_t) + \sum_{t=2}^{T} \log P_{\Delta F0}(f_t - f_{t-1}) + \sum_{t=3}^{T} \log P_{\Delta\Delta F0}(f_t - 2f_{t-1} + f_{t-2}) \right\} \quad (1)$$

3.2 F0 尤度・$\Delta F0$・$\Delta\Delta F0$ の設計

本稿における F0 尤度・$\Delta F0$・$\Delta\Delta F0$ の具体的な設計について述べる．まず F0 尤度には SHS スペクトログラム [9] を各時間フレーム内で正規化したものを用いる．これは計算が容易，高速であるメリットがあり，以下の式で導出される．

$$SHS(t,s) = \sum_{n=1}^{N} (0.84)^{n-1} P(t, s + \log_2 n) \quad (2)$$

$$F0_L(t,s) = \frac{SHS(t,s)}{\sum_{s'=s_{low}}^{s_{up}} SHS(t,s')} \quad (3)$$

ここで s は対数周波数，$P(t,s)$ はスペクトログラムの t 番目の時間フレーム，N は考慮する倍音数，s_{low}，s_{up} は周波数の探索範囲制限幅の下限と上限を表す．本稿では $N = 15$ とした．

$P_{\Delta F0}(f)$，$P_{\Delta\Delta F0}(f)$ はそれぞれ以下のように設計した．

$$P_{\Delta F0}(f) = U(-100, 100) \quad (4)$$

$$P_{\Delta\Delta F0}(f) = \begin{cases} N(f|0, 50^2) & (-50 < f < 50) \\ 0 & (elsewise) \end{cases} \quad (5)$$

f の単位はセントであり，時間フレーム幅は 10 [msec] である．$U(L,U)$ は上限，下限を L，U とする一様分布，$N(f|\mu,\sigma^2)$ は平均，標準偏差を μ，σ とする正規分布を表す．ここで，$P_{\Delta F0}(f)$ を平均 0 の正規分布やラプラス分布とすることも

考えられるが [6]，そうした場合ビブラートなどのピークが平坦に潰れた F0 系列が推定されてしまい，後述する歌い方要素抽出に悪影響を与えるため，本稿では一様分布としている．$P_{\Delta\Delta F0}(f)$ による制約は，F0 推定精度を上げるだけではなく，歌唱 F0 中に必然的に含まれる微細変動 [4] などの歌唱者の意図（歌い方）に関わらない成分を平滑化する効果も期待される．

3.3 ビタビ探索による歌声 F0 推定

式 (1) は 2 重マルコフモデルとなっており，連続した 2 つの時間 $\{t-1,t\}$ における F0 の組み合わせ $\{f_{t-1},t_t\}$ を 1 つの状態とし，ビタビ探索のアルゴリズムを用いて効率的に計算することができる．ビタビ探索は以下の式に従って再帰的に計算する．ここで，$A(\{t-1,t\},\{f_{t-1},f_t\})$, $B(\{t-1,t\},\{f_{t-1},f_t\})$ は累積確率とバックポインタを表している．$A(\{t-1,t\},\{f_{t-1},f_t\})$ は，時刻 $t-1,t$ にそれぞれ F0 が f_{t-1},f_t である確率，$B(\{t-1,t\},\{f_{t-1},f_t\})$ は，時刻 $t-1,t$ にそれぞれ F0 が f_{t-1},f_t があった場合の時刻 $t-2,t-1$ での F0 の値である．

(1) 初期化

$$\forall\{f_1,f_2\},\ A(\{1,2\},\{f_1,f_2\}) = \log P_L(f_1) + \log P_L(f_2) + P_{\Delta F0}(f_2-f_1) \quad (6)$$

(2) 再帰的計算 $(3 \le t \le T)$

$$\begin{aligned} &A(\{t-1,t\},\{f_{-1},f\}) = \\ &\quad \max_{f_{-2},f_{-1}} \Big\{ A(\{t-2,t-1\},\{f_{-2},f_{-1}\}) \\ &\qquad + \log P_L(f) + \log P_{\Delta F0}(f-f_{-1}) \\ &\qquad + \log P_{\Delta\Delta F0}(f-2f_{-1}+f_{-2}) \Big\} \end{aligned} \quad (7)$$

$$\begin{aligned} &B(\{t-1,t\},\{f_{-1},f\}) = \\ &\quad \arg\max_{f_{-2},f_{-1}} \Big\{ A(\{t-2,t-1\},\{f_{-2},f_{-1}\}) \\ &\qquad + \log P_L(f) + \log P_{\Delta F0}(f-f_{-1}) \\ &\qquad + \log P_{\Delta\Delta F0}(f-2f_{-1}+f_{-2}) \Big\} \end{aligned} \quad (8)$$

(3) バックトラック

全ての時間（の組み合わせ）$\{t-1,t\}$ の F0 $\{f_{t-1},t_t\}$ に対してバックポインタ $B(\{t-1,t\},\{f_{t-1},f_t\})$ が計算された．よって，$B(\{t-1,t\},\{f_{t-1},f_t\})$ を後ろ向きにたどることで，式 (1) を最大化する F0 系列 $(\hat{F} := \hat{f}_1,...,\hat{f}_T)$ を得ることができる．

$$\{\hat{f}_{T-1},\hat{f}_T\} = \arg\max_{f_{T-1},f_T} A(\{T-1,T\},\{f_{T-1},f_T\}) \quad (9)$$

$$\hat{f}_t = B(\{t+1,t+2\},\{\hat{f}_{t+1},\hat{f}_{t+2}\})[1], \quad T-2 \ge t \ge 1 \quad (10)$$

ただし，$B(\cdot)[1]$ は $B(\cdot)$ の 1 つめの要素を表す．

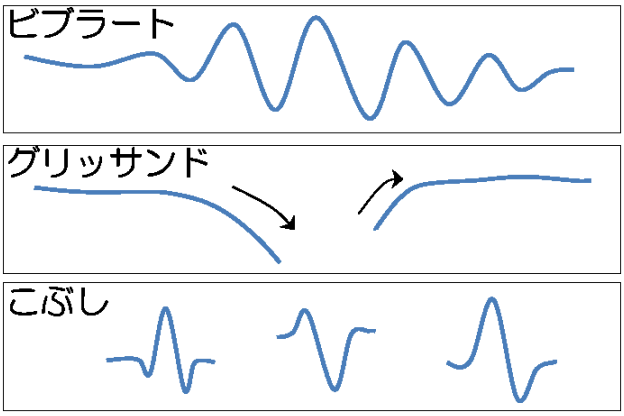


図 2 歌い方要素

4. 歌い方要素の抽出

本章では各歌い方要素を抽出する手法について述べる．まず，二乗誤差最小化に基づくビタビ探索によって，入力音高列と推定 F0 系列の時間的アライメントを計算する．このビタビ探索は，無声区間で必ず次の音高へ移るという制約付きである．歌い方要素抽出は，アライメント結果の各音高について個別に行われる．続く節で，各歌い方要素の抽出を具体的に述べる．

4.1 ビブラート

ビブラートの抽出は中野ら [10] の手法を参考にした．中野らの手法ではビブラートの振幅・周波数の範囲をそれぞれ 30 ～ 150 [cent]，5 ～ 8 [Hz] としていたが，事前に演歌などに現れるビブラートを観測したところ，ビブラートの振幅はその上限を遥かに超え，周波数も下限を下回るケースが多くみられた．そのため，本稿では，ビブラートの振幅に上限を設けず，また周波数範囲は 3 ～ 8 [Hz] としている．

4.2 グリッサンド

歌唱におけるグリッサンドには，グリスダウンとグリスアップがあり，それぞれフレーズ終りに滑らかに音を落とす歌唱法，フレーズ始まりに滑らかに音を上げていく歌唱法を表す (Figure 2)．ある無声区間が T 秒以上の場合，そこでフレーズが途切れているとし，その前後の音高をフレーズ終り・フレーズ始まりとする．

以下のルールに当てはまる区間をグリスダウン（グリスアップ）として抽出する．

（1）ビブラート区間と被らない．

（2）フレーズ終り（始まり）の 後尾（先頭）における，F [cent] 以上の単調減少（増加）．

本稿では，$T = 0.3$，$F = 200$ とした．

4.3 こぶし

こぶし（小節）は演歌や民謡に代表的に現れる，旋律の

装飾的な歌唱法である (Figure 2)．本稿では，こぶしを以下のように特徴を持つ F0 系列上の変化パターンとして抽出する．

(1) ビブラート区間と被らない．
(2) 振幅が $F2$ [cent] 以上の大きなピーク（メインピーク）を 1 つ持つ．
(3) メインピークの前後にそれぞれ 1 つ以下のピーク（サブピーク）を持つ．

ここでピークとは，F0 系列上で極値を持つ点で，且つ前後のピークもしくは立ち上がり点からの変化率が V [cent / sec] を超える点を指す．本稿では $F2 = 150$，$V = 1000$ とした．

5. 歌い方要素のパラメータ化

本章では，前章で抽出した各歌い方要素のパラメータ化について述べる．

5.1 音符情報

同じ歌手でも振幅の違うビブラートなどが観測される．これは音符の並びや音長によって変化するものであると考えられる．そこで次節以降で述べる歌い方要素とともに，本稿では音符の情報として以下の値を保存する．

- 音高
- 音長
- 前後の音高
- 音符の位置 $\in$ { フレーズ始まり, フレーズ終わり, フレーズ中 }

ここで，音長は，前述のアライメントによって音高に割り当てられた時間のうち F0 の存在する（有声）区間の長さである．また，フレーズ中は，その音高の前後に別の音高が繋がっている状態を言う．

5.2 ビブラート

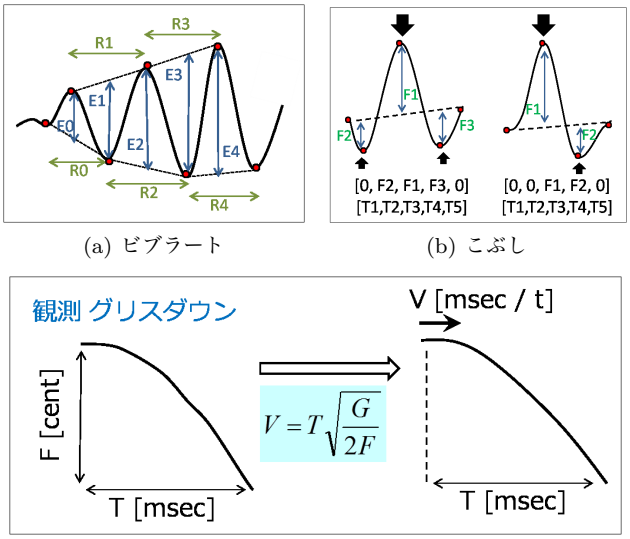
ビブラートは振幅と周波数により，特徴付けられる [11]．ビブラート区間のピーク点（零交差点）を求め，各インデクス（時間）と振幅と周波数を保存する（Figure 3(a)）．ただし，i 番目のピークにおける時刻を t_i [sec]，その時点の（対数）F0 を f_i [cent] としたとき，振幅 E_i と周波数 R_i は以下の式で求められる．

$$E_i = |(f_{i+1} - f_{i-1})\frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}} + (f_{i-1} - f_i)| \quad (11)$$

$$R_i = \frac{1}{t_{i+1} - t_{i-1}} \quad (12)$$

5.3 グリッサンド

グリッサンドは，自由落下パラメータとして保存する (Figure 3(c))．観測されたグリスダウンの横距離（時間幅）T [sec] と縦距離（対数周波数幅）F [cent] から，初速度 V [msec / t] が以下の式で計算される．



(a) ビブラート　(b) こぶし

(c) グリッサンド

図 3　歌い方要素のパラメータ化

$$V = T\sqrt{\frac{G}{2F}} \quad (13)$$

ここで，G [cent / t^2] は重力加速度であり，本稿では $G = 800$ とした．V，T をパラメータとして保存する．グリスアップについても，左右を反転しグリスダウンと同様に考える．

5.4 こぶし

こぶしはメインピークの前後両側にピークを持つもの，片方にのみ持つもの，持たないものが検出される．これらを同じ枠組みで扱うため，各ピークのインデクス（時間）と振幅を，メインピークを中心とする長さ 5 のベクトルとして保存する（Figure 3(b)）．ベクトルの要素は順に，1:始点，2:左のサブピーク，3:メインピーク，4:右のサブピーク，5:終点，における値を保存する．ただし，始・終点における振幅は 0 とし，存在しないサブピークの振幅も 0 となりインデクスは始点または終点と同じになる．こぶしにおける i 番目のピークの大きさ P_i は，その時点の時刻，（対数）F0 をそれぞれ t_i [sec]，f_i [cent] としたとき，以下の式で計算される．

$$P_i = f_i - (\frac{f_5 - f_1}{t_5 - t_1}(t_i - t_1) + f_1) \quad (14)$$

6. 評価実験

6.1 実験条件

実験に用いる楽曲は全て 16 kHz / 16 bits でサンプリングされ，定 Q 変換によって（対数）周波数領域へと変換される．定 Q 変換の時間分解能と周波数分解能はそれぞれ 10 [msec]，6 [cent] とし，周波数範囲は 60 – 6000 [Hz] とした．また Q 値は $(1/(2^{0.01} - 1))/5$ と設定した．前処理として行われる歌声分離は曲全体に対してかけられる．

6.2 研究用データベースを用いた定性的評価

RMC Music Database [12] のポピュラー楽曲 96 曲 (No.3,

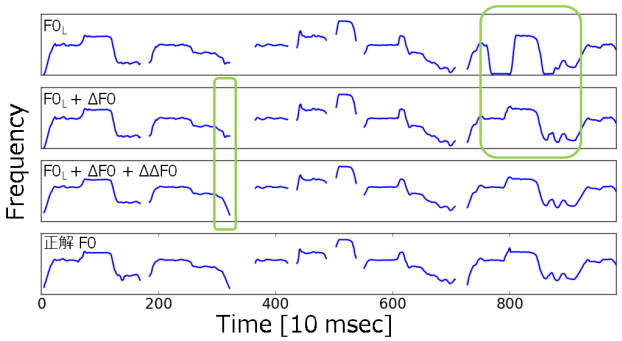


図 4　歌唱 F0 推定の例．上から F0 尤度のみ，F0 尤度と $\Delta F0$，F0 尤度と $\Delta F0$ と $\Delta\Delta F0$（提案法）

表 1　歌唱 F0 推定精度

許容誤差 [cent]	$F0_L$	$F0_L + \Delta$	$F0_L + \Delta + \Delta\Delta$
50	88.59	88.64	**88.82**
25	80.24	80.30	**80.81**

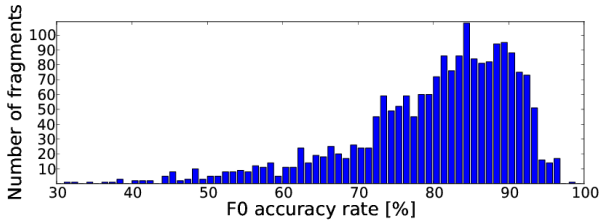


図 5　F0 推定精度分布

4，5，73 を除く）を用いて，歌唱 F0 推定精度を調べた．各曲は時間的に同期した SMF [13] を用いて，10 – 20 音高を含む素片へと分割され，各素片について処理が行われる．生成された素片は計 2001 個である．正解データには手作業で歌唱メロディをアノテーションしたもの [13] を使用した．

F0 推定精度は，マルコフモデルの定式化において F0 尤度のみ，F0 尤度と $\Delta F0$，F0 尤度と $\Delta F0$ と $\Delta\Delta F0$（提案法）を用いたものを比較し，推定誤差は 正解データから 50，25 [cent] としたものをそれぞれ計った．全素片の平均推定精度を表 1 に示す．$\Delta F0$，$\Delta\Delta F0$ の制約によって大きな推定精度の向上は見られないが，いくつかの素片で，フレーズ終わりなどの歌唱音量が小さくなっている箇所において制約による推定誤りの改善が見られた (図 4)．また素片全体において，$\Delta\Delta F0$ の制約により歌い方に関わらない微細変動やノイズの平滑化が観測された．これらは，歌い方要素抽出の精度向上に貢献していると考えられる．

図 5 は，提案法による F0 推定精度 (許容誤差 25 [cent]) の素片数のヒストグラムである．25 [cent] という比較的小さい許容誤差であっても，多くの素片において 80 % を超える高精度な F0 推定を実現できている．60 % を下回るような低精度な素片も一定数存在するが，これらの多くはユニゾン歌唱であったり伴奏音に比べて明らかに歌唱音量が小さいなど，歌い方要素を抽出に適さないものが多かった．ユーザが分析する素片を選択する際には，そのようなものが除かれることが想定されるため，高い F0 推定精度が期待される．

なお，歌い方要素抽出という目的において，本手法の F0 推定精度が十分であるとは言えない．本手法で F0 推定は，音高列のアライメントの前処理にもなっており，実際 F0 推定誤りにより音高列アライメントにも誤りが生じる．また，F0 推定の制約には音高の最大，最小のものしか使用されていない．音高列アライメントが既知であれば，全ての音高を考慮した F0 推定が行えるが，F0 推定がなされていないと音高列アライメントはなされない．この卵と鶏の問題を解決する，F0 と音高列アライメントを同時に推定するモデルを作ることによって，さらに高精度な F0 推定と音高列アライメントを実現できると考えている．

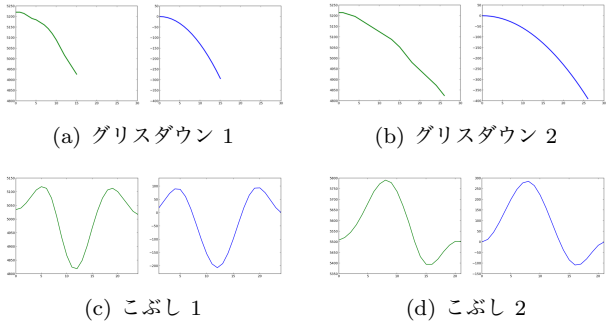


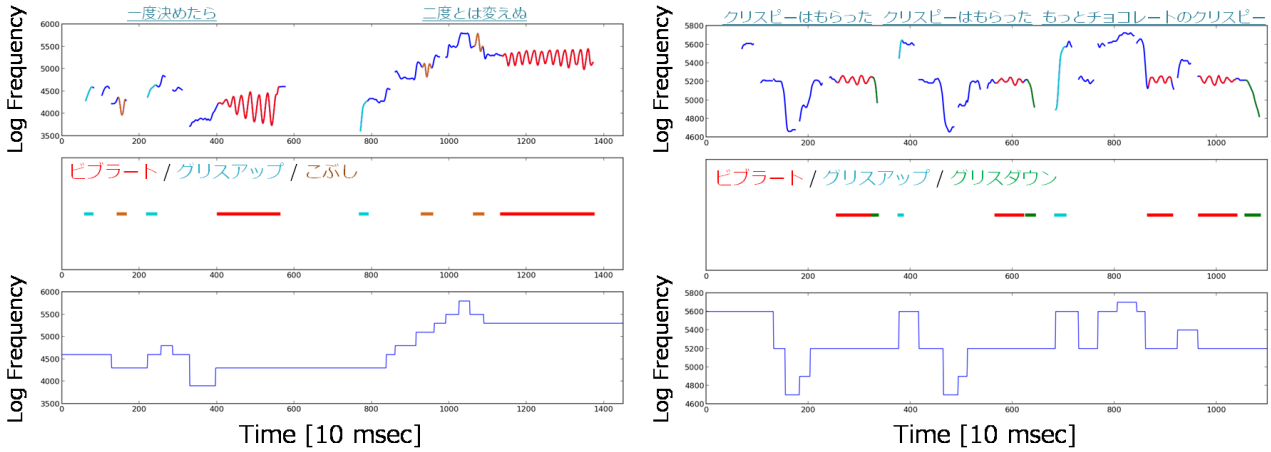
図 7　歌い方要素の再合成．左（緑）が生データ，右（青）が保存パラメータから再合成した歌い方要素．

6.3　市販楽曲からの歌い方要素抽出

市販楽曲 2 曲を用いて本手法の歌い方要素抽出の動作を確認した．用いた楽曲は『人生一路（美空ひばり）』の A メロ部と，『クリスピー（スピッツ）』のサビ部である．前者は，日本の伝統的な歌謡である「演歌」であり，大きなビブラートやこぶしが頻繁に現れる歌唱法が特徴である．後者は，日本のポップス曲であるが，特にこのボーカルの歌唱法の特徴としてグリスダウンを多用することが挙げられる．

図 6 に実験結果を示す．上から順に，推定された F0，検出された歌い方要素，音高列と推定 F0 のアライメント結果である．前者の楽曲（図 6(a)）では，演歌に特徴的に現れる振幅の大きく周期の大きいビブラートやこぶしが検出されている．また，グリスアップは演歌の力んだ歌い方に付随していると考えられる．後者の楽曲（図 6(b)）では，フレーズ終わりにおける頻繁なグリスダウンが検出されている．

図 7 に上記で抽出したパラメータから歌い方要素を再合成した結果を示す．グリスダウンは『クリスピー』の 2，3 つ目、こぶしは『人生一路』の 2，3 つ目のものを取り上げている．また，こぶしの再合成はスプライン補間によって行われている．グリスダウン，こぶしとも，大きさや形状の違うものが同一保存形式のパラメータから再合成され，生データのおおよその形状を保持できていることが確認できる．



(a) 『人生一路（美空ひばり）』　(b) 『クリスピー（スピッツ）』

図 6　市販楽曲からの歌い方要素抽出

7. おわりに

本稿では，伴奏付き歌唱からビブラートなどの歌い方要素を抽出する手法について述べた．本手法では歌声分離の後，歌唱音高列による制約とビタビ探索を用いて F0 推定を高精度に行う．推定 F0 からビブラート・グリッサンド・こぶしを抽出し，設計されたモデルに従ってパラメータとして保存される．実験では市販楽曲から本手法により歌い方要素を抽出し，保存パラメータからの再合成も行えることを確認した．今後は，それぞれの要素がどの程度歌手の歌い方の個人性に関係しているのかを調べるとともに，F0 上の変動に限らない他の要素の模索・導入も検討していく必要がある．

参考文献

[1] Nakano, T. and Goto, M.: VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation, *Proc. SMC*, pp. 343–348 (2009).
[2] Morise, M., Onishi, M., Kawahara, H. and Katayose, H.: v.morish 09: A Morphing-Based Singing Design Interface for Vocal Melodies, *Proc. ICEC*, Vol. 5709, pp. 185–190 (online), DOI: 10.1007/978-3-642-04052-8_18 (2009).
[3] Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y. and Tokuda, K.: Recent Development of the HMM-based Singing Voice Synthesis System - Sinsy, *Proc. ISCA Tutorial and Research Workshop on Speech Synthesis*, pp. 211–216 (2010).
[4] Ohishi, Y., Kameoka, H., Mochihashi, D. and Kashino, K.: A Stochastic Model of Singing Voice F0 Contours for Characterizing Expressive Dynamic Components, *Proc. Interspeech* (2012).
[5] Goto, M.: PreFEst: A Predominant-F0 Estimation Method for Polyphonic Musical Audio Signals, *Proc. MIREX* (2005).
[6] 藤原弘将，後藤真孝，奥乃　博：歌声の統計的モデル化とビタビ探索を用いた多重奏中のボーカルパートに対する音高推定手法，情報処理学会論文誌，Vol. 49, No. 10, pp. 3682–3693 (2008).
[7] Salamon, J. and Gmez, E.: Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics, *IEEE TASLP*, Vol. 20, No. 6, pp. 1759–1770 (2012).
[8] Rafii, Z. and Pardo, B.: Music/Voice Separation using the Similarity Matrix, *Proc. ISMIR*, pp. 583–588 (2012).
[9] Hermes, D. J.: Measurement of pitch by subharmonic summation, *J. Acoust. Soc. Am.*, Vol. 83, No. 1, pp. 257–264 (online), DOI: 10.1121/1.396427 (1988).
[10] Nakano, T. and Goto, M.: An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features, *Proc. Interspeech* (2006).
[11] Migita, N., Morise, M. and Nishiura, T.: A study of vibrato features to control singing voices, *ICA2010* (2010).
[12] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp. 287–288 (2002).
[13] Goto, M.: AIST Annotation for the RWC Music Database, *Proc. ISMIR*, pp. 359–360 (2006).