情報処理学会論文誌 Vol. 49 No. 11 3789–3797 (Nov. 2008)

相平面に描かれる歌声の基本周波数軌跡：歌唱者の意図する音高目標値系列の推定とハミング検索への応用

大石康智[†1] 後藤真孝[†2]
伊藤克亘[†3] 武田一哉[†1]

本論文では，歌声の基本周波数（F0）軌跡に含まれる歌唱者の意図する音高目標値と歌声の動的変動成分を可視化することのできる新しい表現方法を提案する．F0 軌跡が自励系の微分方程式に従って生成されるものと想定し，その解（F0 軌跡）の性質を調べるために，F0 とその時間微分からなる 2 次元平面（相平面）上に F0 軌跡を表現する．この相平面にはいくつもの渦軌跡（アトラクタ）が描かれ，これらの渦の中心は歌唱者の意図する音高目標値に相当する．一方で，この渦の中心にいたるまでの曲線の軌跡に，歌声の動的変動成分が表現される．そこで，この相平面におけるアトラクタを確率的にモデル化し，その中心を追跡することによって，F0 軌跡から歌唱者が本来歌おうとする音高目標値の時系列を推定する手法を提案する．推定結果をハミング検索における旋律の類似尺度に適用したところ，従来の F0 軌跡の DP マッチングによる照合と同等以上の結果が得られたことから，提案手法の有効性を確認できた．

Sung Melodic Contour Characterized in Phase Plane: Estimation of Target Note Sequence from the Melodic Contour and Its Application for Query-by-humming

Yasunori Ohishi,[†1] Masataka Goto,[†2]
Katunobu Itou[†3] and Kazuya Takeda[†1]

In this paper, we propose a new visualization technique of a sung melodic contour, which can characterize both musical-note information and the dynamics of singing behaviors included in the melodic contour. We assume that the fundamental frequency (F0) trajectories are generated by a dynamic system and represented in a two-dimensional phase plane which consists of F0 and its differential. In this plane, a fluctuation in a sung melody can be modeled by a damped oscillation of the dynamic system and appears as a curling trajectory around a certain target point, i.e., an attractor of the system. The advantage of this modeling is that the location of each attractor corresponds to the F0 of its target musical note and typical singing behaviors can be characterized by the shape of curling trajectories. By modeling this representation stochastically and tracking locations of attractors, we propose an estimation method of the target note sequence from an F0 contour and define a melodic similarity measure for query-by-humming (QBH) applications. Experimental results show that our QBH method with the proposed similarity measure is superior to a conventional dynamic-programming-based QBH method.

1. はじめに

本研究では，歌声の F0 軌跡から歌唱者の意図する旋律概形や，ビブラートやオーバシュートのような歌声特有の動的変動成分を特徴付ける信号モデルの構築を目指す．歌声は，多くのジャンルの音楽を特徴付ける重要な要素の 1 つであり，現在様々な研究がされている[1)–3)]．しかし，歌声の動的変動成分やそれに基づく歌唱スタイルのモデル化についてはまだ十分に検討されていない．またハミング検索では，歌唱された歌声の F0 軌跡から，旋律を構成する音高列を正しく推定して，楽曲データベース中の旋律と照合する必要がある．従来は，音声信号のパワーを利用して，F0 軌跡を音高と音長を表すシンボル列に変換し，*n*-gram モデルのような離散的な確率表現を利用することが一般的であった[4)–8)]．しかし，タタタやチャチャチャで歌うハミングに比べて，歌詞付きの歌声は，発声する音素の種類が多いため，発声における声帯振動の影響を受けて，F0 軌跡に微細な変動成分が多く含まれる．さらにビブラートやオーバシュートのような動的変動成分をも含むため，旋律を正しくシンボル列で表現することが難しい．F0 軌跡そのものを DP マッチングによって照合して検索する方法が提案されている[9)–11)] が，それでも歌声の動的変動成分の影響を受けて検索性能が低下してしまう．**図 1** は，楽譜に記述される旋律の音高列とその旋律を歌詞付きで歌唱したときの F0 軌跡を図示した例である．F0 軌跡には，目的音高に達するまでの立ち上がりや目的

†1 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†2 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
†3 法政大学情報科学部
Faculty of Computer and Information Science, Hosei University

音高より大きく振れてしまうオーバシュート，音高が安定するときに振動させるビブラートなどの動的変動成分が観測される．我々は，このような F0 軌跡から旋律を構成する音高列を正しく推定するためには，動的変動成分そのものを適切にモデル化する必要があると考えている．

また，隠れマルコフモデル（HMM）や制動 2 次系のインパルス応答を利用した F0 制御モデルによって，自然性かつ明瞭性のある歌声合成が実現された[12),13)]．ただ，同じ旋律でも人それぞれ歌い方が異なるように，合成音声の多様な歌唱スタイルについてはさらなる検討課題である．そのためにも歌声の F0 軌跡に含まれる動的変動成分を特徴付けるモデルが求められている．

本論文では，歌声の F0 軌跡から歌唱者の意図する音高目標値と動的変動成分を可視化するための新しい表現方法を提案する．F0 軌跡が自励系の微分方程式に従って生成されると想定し，F0 とその時間微分で構成される相平面上に F0 軌跡を表現する．このとき，音高目標値は相平面に現れる渦の中心，動的変動成分は渦の中心にいたるまでの曲線の軌跡によって表現される．さらにこの表現方法を利用して，F0 軌跡から音高目標値の時系列を推定する手法を提案する．提案手法を旋律の類似尺度に適用し，ハミング検索実験を行ったところ，従来の F0 軌跡の DP マッチングによる照合よりも検索結果を適切に絞り込むことが可能であることを確認した．

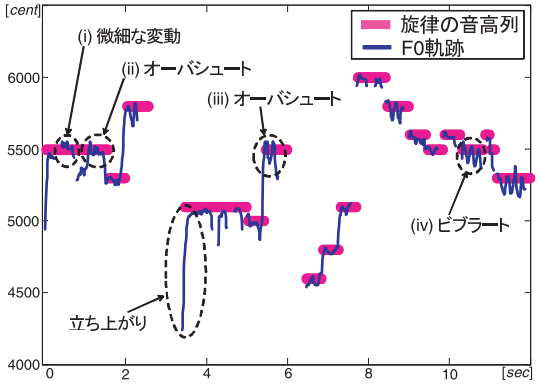

図 1　旋律の音高列と，それを歌詞付きで歌唱した歌声の F0 軌跡：F0 軌跡には，音高目標値に対する立ち上がりやオーバシュート，ビブラートなどの動的変動成分が観測される

Fig. 1　Note sequence of a melody and the F0 contour estimated from a voice signal singing its melody with lyrics: the dynamics of singing behaviors are observed in the F0 contour.

以下，2 章では，相平面を利用した F0 軌跡の表現方法について述べ，3 章では相平面の F0 軌跡から歌唱者の意図する音高目標値系列を推定する手法を提案する．4 章では提案手法によって推定された時系列をハミング検索のための旋律の類似尺度に適用し，その有効性を確認するための評価実験を行う．5 章では実験結果を考察し，6 章でまとめと今後の課題について述べる．

2. 相平面における歌声の F0 軌跡

図 2 (a) は，F0 を表す y とその時間微分 $\dot{y}$ からなる 2 次元平面（相平面）に描かれる図 1 の歌声の F0 軌跡である．ここで時刻 n の F0 を y_n と表す．F0 は，de Cheveigné らの提案した YIN [14)] を利用して 10 ms ごとに推定される．なお，Hz で表された周波数 y_{Hz} を，次のように，cent で表された対数スケールの周波数 y_{cent} に変換する．

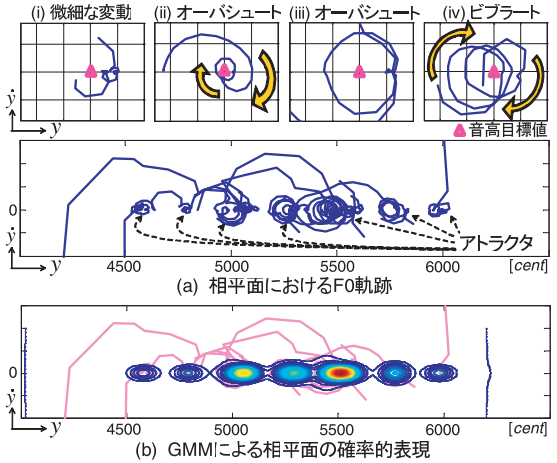

図 2　相平面（F0 とその時間微分からなる 2 次元平面）に描かれる歌声の F0 軌跡：(a) は図 1 の F0 軌跡を相平面に図示した結果である．音高遷移が，複数のアトラクタとそれらを遷移する動きによって表現される．オーバシュートは螺旋，ビブラートは楕円を描く軌跡によって表現される．(b) は相平面における F0 軌跡の分布を GMM によって学習した結果である

Fig. 2　Sung F0 contour mapped onto a phase plane (two-dimensional plane consisting of F0 and its differential): (a) shows the estimated F0 contour of Fig. 1 on the phase plane. An overshoot after a note change and a vibrato within a musical note appear as a spiral pattern and an ellipsoidal pattern, respectively. By fitting Gaussian mixture models to F0 trajectories on the phase plane, stochastic representation of the F0 dynamics (b) can be constructed.

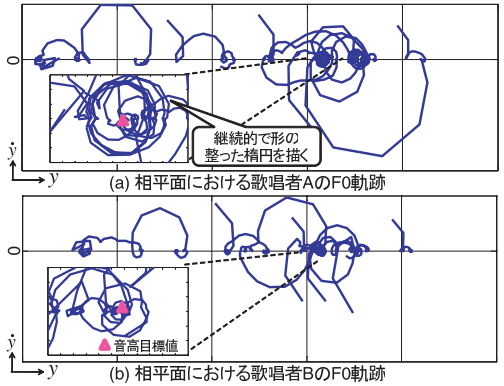

図 3　2 人の異なる歌唱者による同じ旋律の歌声の F0 軌跡：音高目標値を中心に振動するビブラート（アトラクタの様相）や音高の遷移（アトラクタ間の遷移）が歌唱者ごとに異なる

Fig. 3　F0 contours of the same melody sung by two different singers: the degree of vibrato and continuous transitions between notes vary among singers.

$$y_{\mathrm{cent}} = 1200 \log_2 \frac{y_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}} \tag{1}$$

一方，$\dot{y}_n$ は時刻 n の F0 の時間微分を表し，以下のように微小区間（50 ms）の F0 の回帰係数 ΔF0 で近似する．

$$\dot{y}_n \simeq \Delta\mathrm{F0} = \frac{\sum_{k=-2}^{k=2} k \cdot y_{n+k}}{\sum_{k=-2}^{k=2} k^2} \tag{2}$$

以上で求めた y_n と $\dot{y}_n$ を時刻 n の観測ベクトル $\mathbf{y}_n$ と表現する．

我々は，歌声の F0 軌跡が自励系の微分方程式に従って生成されるものと想定し，その解（F0 軌跡）の性質を新たな視点で眺めることのできる相平面を利用する．この平面では，解曲線が渦を描きながら，ある点に引き寄せられる動き（アトラクタ）が観測される．また，アトラクタから別のアトラクタに遷移する動きが観測される．これらのアトラクタの中心は，歌唱者が意図する音高目標値に相当する．一方，アトラクタの中心にいたるまでの渦軌跡には，歌声の動的変動成分が表現される．たとえば，音高安定時に準周期的な振動を繰り返すビブラートは，音高目標値を中心に楕円を描く軌跡として観測される．また，音高が遷移するときに目的音高より大きく振れてしまうオーバシュートは，螺旋を描きながらアトラクタに引き寄せられる軌跡として観測される（図 2 (a)）．以上のように F0 軌跡を相平面上に描くことによって，歌唱者の意図する音高目標値と動的変動成分を可視化できる．

さらに同じ旋律を 2 人の異なる歌唱者が歌ったときの F0 軌跡を**図 3** に示す．歌唱者 A の F0 軌跡には，歌唱者 B に比べて，継続的に形の整った楕円軌跡が観測される．このことから歌唱者 A は，歌唱者 B よりも振幅の大きいビブラートをかける傾向があるといえる．また，ビブラートだけでなく，音高の遷移の仕方にも歌唱者間で違いがみられる．以上のように提案手法は，歌唱者の歌い方や個性をも可視化できるため，様々な応用（たとえば，歌唱者識別，歌唱力評価，歌唱スタイルの転写など）に対して重要な基礎技術であると考えられる．

3. 相平面を利用した歌声の音高目標値の推定

相平面を利用して，観測される F0 軌跡 $y_1, \ldots, y_N$ から歌声の動的変動成分を除去し，歌唱者の意図する音高目標値の時系列 $m_1, \ldots, m_N$ を推定する．図 2 (a) の (i)〜(iv) は，すべて同じ音高目標値の周りで観測される動的変動成分である（図 1 の (i)〜(iv) に対応する）．同じ音高目標値であっても，それ以前の音の並びや歌唱者の歌い方の影響を受けて，その周辺で観測される動的変動成分には，ばらつきが生じる．本論文では，この音高目標値の周りの動的変動成分が，確率的に変動するものであると想定する．このばらつきを吸収し，歌唱者の意図する音高目標値を推定するために，アトラクタを確率分布として表現する．図 2 (b) に示すように，相平面の F0 軌跡の分布を混合ガウス分布（GMM）によって確率的に表現すると，確率密度の極大値がアトラクタの中心に対応することが分かる．ここで GMM の混合数は 16 とした．

そこで**図 4** に示すように，F0 軌跡をフレーム化処理し，フレームごとに相平面における F0 軌跡の分布を GMM によって学習して，アトラクタの中心を追跡する．まず，前処理として，無声音や休符のため F0 が推定されない区間，またそれによって $\dot{y}_n$ を計算できない区間はすべて除去する[⋆1]．フレーム長を $2T+1$ として，時刻 $n-T$ から時刻 $n+T$ の観測ベクトルの集合 $\mathcal{Y}^{(n)}$ をフレーム n と定義する．このフレーム n において推定される GMM のパラメータを $\boldsymbol{\Theta}^{(n)} = \{w_1^{(n)}, \ldots, w_M^{(n)}, \boldsymbol{\mu}_1^{(n)}, \ldots, \boldsymbol{\mu}_M^{(n)}, \boldsymbol{\Sigma}_1^{(n)}, \ldots, \boldsymbol{\Sigma}_M^{(n)}\}$ とする．ここで，M は GMM の混合数であり，各ガウス分布の重み，平均，分散は，それぞれ w，$\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$ と表す．このとき，$\mathcal{Y}^{(n)}$ の尤度 $p(\mathcal{Y}^{(n)}|\boldsymbol{\Theta}^{(n)})$ を最大にする $\hat{\boldsymbol{\Theta}}^{(n)}$ を，EM アルゴリ

⋆1 F0 が推定されない区間を除去すると，分析フレーム中に不連続な境界を含む場合がある．このような境界では，本論文で定義する $\dot{y}_n$ を計算できないため，F0 が推定されない区間とともに，それによって $\dot{y}_n$ を計算できない区間（境界の前後 3 点，つまり 60 ms）を除去した．

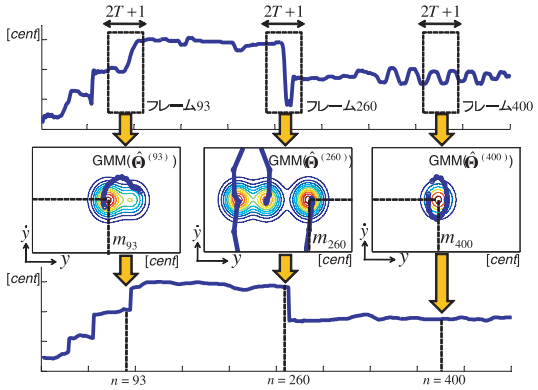

図 4　歌唱者の意図する音高目標値の推定手法：フレームごとに F0 軌跡を相平面上に描き，その分布を GMM で学習する．確率密度が最大となる F0 の値 m_n をフレーム n における音高目標値とする

Fig. 4　A method of estimating the target note sequence from the F0 contour: for each frame, the F0 contour is mapped onto the phase plane and the distribution is modeled by GMM. We estimate F0 value m_n of which the probability density is the maximum as the target note in the frame n.

ズムによって推定する．ここで，推定する $\mathbf{\Sigma}$ は対角共分散行列とした．最後に，$\hat{\mathbf{\Theta}}^{(n)}$ による GMM の確率密度が最大となる F0 の値を，フレーム n の観測ベクトルによって描かれるアトラクタの中心 m_n とする．GMM の確率密度の最大値は解析的に求めることが難しいので，相平面を格子に区切り，数値的に求める．

図 4 の下図は，提案手法によって推定された音高目標値の時系列を示す．$T = 100\,\mathrm{ms}$, $M = 4$ とした場合である．ビブラートやオーバシュートのような動的変動成分が除去され，歌唱者が本来歌おうとする音高目標値の時系列が推定される．この推定結果は，たとえば，ハミング検索における旋律の類似尺度に利用できる．動的変動成分の影響を受けて従来の F0 軌跡の DP マッチングでは正しく照合できない歌声に対して，その距離の改善が期待される．

4. 評 価 実 験

F0 軌跡から歌唱者の意図する音高目標値系列を推定し，その結果を旋律の類似尺度に利用したハミング検索実験を行う．歌唱者の歌声（以後，入力信号と呼ぶ）と楽曲データベースの各楽曲の旋律（以後，参照信号と呼ぶ）の F0 軌跡を提案手法によって音高目標値の時系列に変換し，DP マッチングによって時系列間の距離を求めて検索結果を算出する．また従来法[9)–11)] として，F0 軌跡間の DP マッチングによる距離に基づく検索結果も算出する．前処理として，各信号ごとに推定される F0 軌跡の平均値を計算し，F0 軌跡からこの平均値を減算する．これは，歌唱者が原曲の旋律とは異なる音の高さで歌う移調に対応するためである．

DP マッチングは，入力信号の系列 I_t（$1 \le t \le T$）と参照信号の系列 R_s（$1 \le s \le S$）との最適マッチング系列を求める手法である．以下のように局所的な傾斜を 1/2 と 2 の間に制限した漸化式を利用して系列間の距離を求める．

$$g(I_t, R_s) = \min \begin{bmatrix} g(I_{t-2}, R_{s-1}) + 2d(I_{t-1}, R_s) + d(I_t, R_s) \\ g(I_{t-1}, R_{s-1}) + 2d(I_t, R_s) \\ g(I_{t-1}, R_{s-2}) + 2d(I_t, R_{s-1}) + d(I_t, R_s) \end{bmatrix} \tag{3}$$

ここで，$g(I_1, R_1) = 2d(I_1, R_1)$ として計算を繰り返し，最後に $D = g(I_T, R_S)/(T+S)$ として，2 つの時系列間の時間正規化後の距離が求まる．局所距離は，$d(I_t, R_s) = |I_t - R_s|$ とする．

4.1　楽曲データベース

「RWC 研究用音楽データベース：ポピュラー音楽」（RWC-MDB-P-2001）[15)] の計 100 曲から，歌唱の出だしの部分（以後，出だしと呼ぶ）と盛り上がる主題の部分（以後，サビと呼ぶ）の 2 カ所を切り出し，全 200 種類の参照信号からなる楽曲データベースを構築した．これらの信号の切り出し区間は，その部分の歌詞の始まりから区切りの良いところまでとし，平均 11.7 秒であった．本来ならばこれらの信号から F0 を推定すべきだが，今回は提案手法の性能の上限を調べるために，楽曲データベースに関しては F0 を手作業でラベル付けした結果[16)] を用いた．

4.2　歌声データベース

歌声研究用音楽データベース「AIST ハミングデータベース」[17)] の一部である，日本人歌唱者 75 人（男性 37 人，女性 38 人）が，上記の 200 種類の参照信号のうち 50 種類を歌詞付きで歌唱した計 3,750 サンプルを入力信号として利用する．歌唱者は伴奏なしで，自由なテンポで歌唱した．歌唱時間は，平均 12.0 秒であった．歌唱者は，初めて聴くポピュラー音楽をうろ覚えの状態で歌唱したため，収録された歌声は原曲の旋律に比べて多少の揺れを含んでいる．これらの揺れは，ビブラートのような動的変動成分によるものばかりでなく，うろ覚えのために生じた音符の挿入や置換，削除によるものでもある．したがって，計

3,750 サンプルから比較的正しく歌えている入力信号を手作業で選定し，その結果，50 種類の旋律を歌唱した 2,073 サンプルを実験で利用する．

4.3 評 価 方 法

評価方法として，以下の 3 つの尺度を利用する．

4.3.1 1 位検索率

入力信号と 200 種類の参照信号との DP マッチングによる距離に基づいて，参照信号をランク付けする．そして，この入力信号の正解に相当する参照信号の順位を求め，これを検索結果とする．2,073 サンプルの入力信号のうち，検索結果が 1 位となった入力信号の割合を 1 位検索率とする．入力信号による検索の正確性を確認するためにこの尺度を利用する．

4.3.2 平均逆順位

前節で求めた各入力信号の検索結果に基づいて，平均逆順位（Mean reciprocal rank, MRR）を計算する．

$$\mathrm{MRR} = \frac{1}{L}\sum_{k=1}^{200}\frac{r_k}{k} \tag{4}$$

ここで，L は入力信号の総数（本実験では 2,073），r_k は，検索結果の順位が k 位となった入力信号の数とする．各入力信号の検索結果の順位が高くなるほど，MRR は大きくなる．入力信号によって 1 位に検索対象の楽曲が検索されなくても，提案手法によってどの程度，検索順位の改善を図ることができたかを確認するためにこの尺度を利用する．

4.3.3 再現率と適合率

入力信号と参照信号との DP マッチングによる距離と閾値 ε に基づいて，以下のように再現率と適合率を計算する．

$$\text{再現率} = \frac{R}{C},\quad \text{適合率} = \frac{R}{N} \tag{5}$$

R：各入力信号とその正解に相当する参照信号との DP マッチングによる距離を計算し，その距離が閾値 ε 以下となった入力信号の数

C：入力信号の総数

N：各入力信号ごとに 200 種類の参照信号と DP マッチングによる距離を計算し，その距離が閾値 ε 以下となった参照信号の数

提案手法によってどの程度，入力信号とその正解に相当する参照信号との DP マッチングによる距離の改善を図ることができたかを確認するためにこの尺度を利用する．

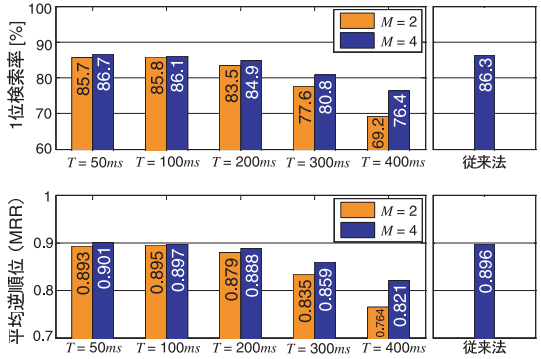

図 5　1 位検索率と平均逆順位による検索性能の評価

Fig. 5　Evaluation of retrieval performance using rank 1 retrieval rate and mean reciprocal rank.

4.4 実 験 結 果

図 5 では，1 位検索率と平均逆順位によって，提案手法の検索性能を評価する．提案手法のフレーム長を決めるパラメータ T を 50，100，200，300，400 ms，GMM の混合数 M を 2，4 と変化させた．T を 50 ms より小さくとすると，フレームによっては安定して GMM を推定することが難しいため，T の下限を 50 ms とした．T を大きくするにつれて性能が低下した．また，すべての T に対して，混合数 M が 2 よりも 4 の方が性能が高い．従来法と比較すると，$T = 50$ ms，$M = 4$ のときに，1 位検索率において 0.4 ポイント，平均逆順位において 0.005 ポイント，性能が改善された．

図 6 では，4.3.3 項の閾値 ε を変化させて再現率と適合率を求め，それらの値によって描かれる曲線から提案手法の検索性能を評価する．曲線が右上方に描かれるほど検索性能が高いことを意味する．図 5 の結果より，提案手法の GMM の混合数は $M = 4$ として，フレーム長を決めるパラメータ T を 50，100，200，300，400 ms と変化させた．図 6 の拡大図から，$T = 50, 100$ ms のときに，従来法を上回る再現率と適合率が得られた．以上，いくつかの観点から従来法との性能比較を行った結果，いずれの観点においても従来法と同等以上の結果が得られることが確認された．

図 7 では，入力信号 2,073 サンプルのうち，歌唱する旋律ごとに 1 位検索率を求め，提案手法によって性能が改善された旋律と性能が低下した旋律を示す．**図 8** では，図 7 で取り上げた旋律の平均逆順位を示す．図 5，図 6 の検索性能から，提案手法の GMM の混合数 M は 4，フレーム長を決めるパラメータ T は 50 ms の場合を取り上げる．「P001 出だ

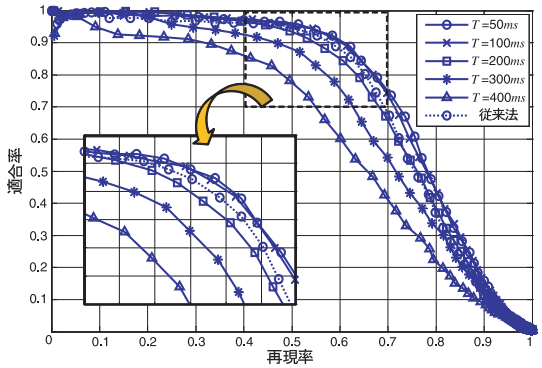

図 6　再現率と適合率による検索性能の評価
Fig. 6　Evaluation of retrieval performance using the recall and precision curve.

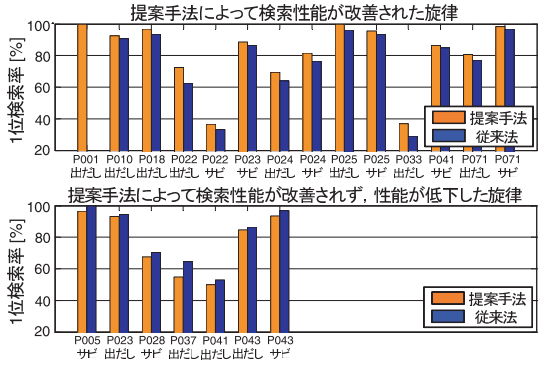

図 7　入力信号の旋律ごとに算出した 1 位検索率
Fig. 7　Rank 1 retrieval rate calculated for each melody of input signals.

し」とは，楽曲番号 P001（RWC-MDB-P-2001 No.1）の出だしの旋律を歌唱した入力信号の集合を意味する．図 7 の上図の「P001 出だし」は，該当する入力信号が 1 つだけであり，従来法では検索結果の順位が 2 位であったため，1 位検索率が 0%となった．図 7，図 8 より，提案手法によって，14 個の旋律は 1 位検索率と平均逆順位がともに改善される結果となったが，7 個の旋律は性能が低下したことが分かる．したがって，提案手法では性能が改善されず，逆に性能が低下してしまう旋律も存在することが分かった．

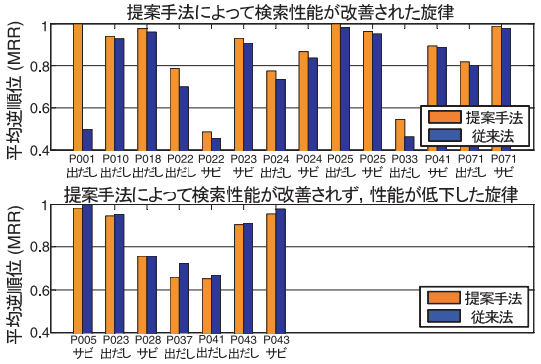

図 8　入力信号の旋律ごとに算出した平均逆順位
Fig. 8　Mean reciprocal rank calculated for each melody of input signals.

5. 考　　察

実験結果をふまえて，提案手法による旋律の類似尺度の有効性について考察する．

5.1 提案手法によって検索性能が改善される場合

まず，検索性能が改善される様子を確認するために，楽曲番号 P041 のサビの旋律を歌唱した，ある入力信号を例にあげて説明する（**図 9**）．従来法では，この入力信号の正解に相当する P041 のサビの参照信号の検索順位は 2 位であり，楽曲番号 P008 の出だしの参照信号が 1 位となった．しかし，提案手法によって，この結果が逆転し，P041 のサビの参照信号が 1 位に検索された．提案手法の GMM の混合数 M は 4，フレーム長を決めるパラメータ T は 50 ms の場合である．図 9 の 1 段目は，P041 のサビ，P008 の出だしの本来の旋律の音高列を示す．つまり，これらは楽譜に相当する．図 9 の 2 段目は，これらの旋律を歌唱したときに観測される F0 軌跡を示す．1 段目に示す階段状の旋律概形が，より連続的に遷移する軌跡となる．従来は，これらの F0 軌跡の DP マッチングによる距離を求め，検索を行っていた．このとき，入力信号は P008 の出だしの参照信号との距離の方が近い結果となった（$D = 68.2$）．一方で，提案手法の，F0 軌跡から歌唱者の意図する音高目標値系列を推定する処理によって，図 9 の 4 段目に示す結果が得られる．F0 軌跡に含まれる動的変動成分が除去されるため，これは F0 軌跡を平滑化する処理ともいえる．これらの時系列間の DP マッチングによる距離を求めた場合，入力信号と P041 のサビの参照信号との距離が改善され（$D = 49.0$），1 位に検索できた（図 9 の 5 段目）．特に，点線で囲んだ部分にお

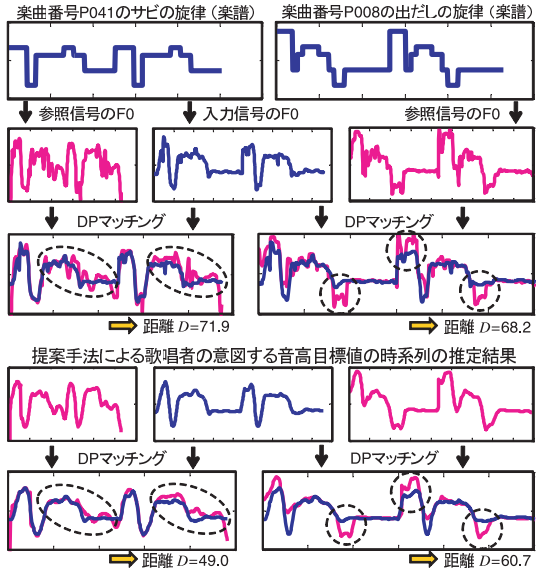

図 9　提案手法によって検索結果の順位が改善された入力信号
Fig. 9　A query whose retrieval performance is improved by the proposed method.

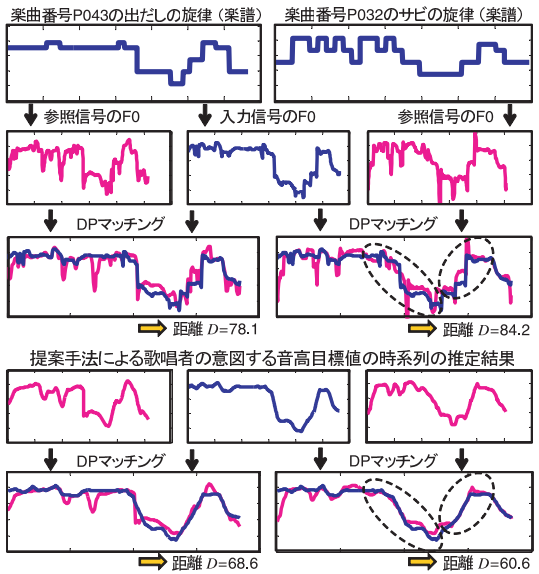

図 10　提案手法によって検索結果の順位が低下した入力信号
Fig. 10　A query whose retrieval performance is degraded by the proposed method.

いて，距離の改善を確認できる．したがって，提案手法の，F0 軌跡を平滑化して歌唱者の意図する旋律概形を推定する処理が，検索性能の改善につながることが分かった．

5.2　提案手法によって検索性能が低下する場合

図 7，図 8 に示すように，歌唱する旋律によっては検索性能が低下する．これは，提案手法による F0 軌跡の過度の平滑化によって，旋律を構成する音高パターンが変形されたことが原因であると考えられる．その例を**図 10** に示す．従来法では，この入力信号の正解にあたる，P043 のサビの参照信号が 1 位に検索されたが，提案手法によって，P032 の出だしの参照信号が 1 位に検索された．図 10 の 4 段目に示すように，提案手法の F0 軌跡の平滑化によって，入力信号が P032 の出だしの参照信号と似た時系列になることが分かる．また図 10 の 5 段目に示すように，P032 の出だしの参照信号の方が DP マッチングによる距離が近くなった．この F0 軌跡の過度の平滑化が検索性能に影響を及ぼすことは，図 5 の T の増加とともに性能が低下することからも確認できる．したがって，提案手法では，短時間に音符を多く含む旋律に対しては，その動きに追随できず，逆に旋律の概形を変形させてしまう可能性があると考えられる．そこで，フレームごとに混合数を変化させて GMM を学習すること，また動的にフレーム長を変化させながら GMM を学習するなどの改善策を検討する必要がある．

5.3　提案手法による F0 軌跡の平滑化作用に関する考察

図 9 では，提案手法による F0 軌跡の平滑化の作用が検索性能の改善につながることを確認した．そこで最後に，入力信号と参照信号の F0 軌跡を移動平均フィルタ（窓長は $2T+1$）によって平滑化し，得られた時系列間の DP マッチングによる距離に基づいて検索性能を算出する．提案手法と単純な F0 軌跡の平滑化に基づく検索性能を比較するためである．**図 11** では，1 位検索率と平均逆順位から検索性能を比較する．提案手法の GMM の混合数 M は 4 とした．$T = 400\,\mathrm{ms}$ における 1 位検索率を除いて，提案手法の方が高い性能が得られた．したがって，単純に F0 軌跡を移動平均で平滑化するよりも，相平面に GMM を配置することによって，F0 軌跡から音高が安定する部分を追跡しながら平滑化する提案手法の有効性を確認できた．

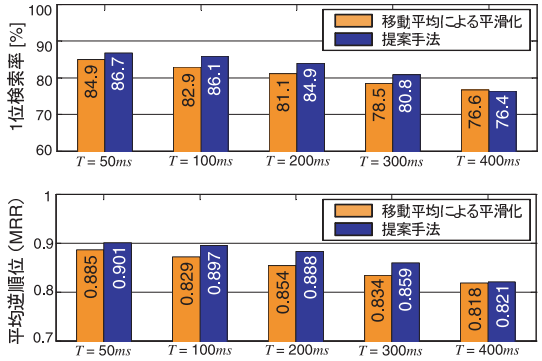

図 11　移動平均フィルタによって平滑化された F0 軌跡を DP マッチングによる照合に利用したときの検索性能と提案手法の性能との比較
Fig. 11　Retrieval performance comparison between the proposed method and Dynamic Programming using F0 contour smoothed by a moving average filter.

6. おわりに

相平面を利用すると，歌声の F0 軌跡に含まれる歌唱者の意図する音高目標値と動的変動成分を可視化することができる．本論文では，この相平面に描かれる F0 軌跡から音高目標値の時系列を推定する手法を提案した．歌声特有の動的変動成分が平滑化されて，歌唱者の歌おうとする旋律概形が推定される．さらにこの推定結果をハミング検索のための旋律の類似尺度に利用したところ，従来の F0 軌跡間の DP マッチングに基づく検索性能を改善することができた．動的変動成分の除去を含め，F0 軌跡を平滑化してから DP マッチングを行うことの有効性を確認した．

今後の課題は，提案手法による F0 軌跡の過度の平滑化を抑え，どのような旋律に対しても検索性能を改善することである．そのために，GMM の混合数や分析フレーム長を最適に決定しながら，音高目標値系列を推定する手法を検討する必要がある．また，原理的に F0 が存在しない区間においても歌唱者は音高目標を意識していると考えられる．このような潜在的な音高目標を利用する方法も，重要な将来課題である．

相平面に描かれる歌声の F0 軌跡の性質を，ハミング検索だけでなく，その他の応用に適用することも検討している．たとえば，歌唱者の演奏表現や癖を特徴付ける F0 軌跡の “動き” を相平面上でモデル化することによって，先行研究とは異なる視点に基づく歌唱力評価や歌唱者の識別手法を提案できると考えられる．歌唱者 A の歌い方，歌唱者 B の歌い方というような多様な歌声合成への応用も考えられる．したがって，相平面におけるアトラクタの渦軌跡の形状をさらに分析し，その動きをモデル化するための技術を検討することも今後の課題である．

謝辞　本研究は日本学術振興会特別研究員（DC2）科研費の補助を受けた．

参 考 文 献

1) 河原英紀，片寄晴弘：高品質音声分析変換合成システム STRAIGHT を用いたスキャット生成研究の提案，情報処理学会論文誌，Vol.43, No.2, pp.208–218 (2002).
2) 藤原弘将，北原鉄朗，後藤真孝，駒谷和範，尾形哲也，奥乃　博：伴奏音抑制と高信頼度フレーム選択に基づく楽曲の歌手名同定手法，情報処理学会論文誌，Vol.47, No.6, pp.1831–1843 (2006).
3) 中野倫靖，後藤真孝，平賀　譲：楽譜情報を用いない歌唱力自動評価手法，情報処理学会論文誌，Vol.48, No.1, pp.227–236 (2007).
4) 園田智也，後藤真孝，村岡洋一：WWW 上での歌声による曲検索システム，電子情報通信学会論文誌 D-II，Vol.J82-D-II, No.4, pp.721–731 (1999).
5) Dannenberg, R.B., Birmingham, W.P., et al.: A Comparative Evaluation of Search Techniques for Query-by-Humming Using the MUSART Testbed, *Journal of the American Society for Information Science and Technology*, Vol.58, No.5, pp.687–701 (2007).
6) Song, J., Bae, S.Y. and Yoon, K.: Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System, *Proc. ISMIR 2002* (2002).
7) Pauws, S.: CubyHum: A fully operational query by humming system, *Proc. ISMIR 2002* (2002).
8) Pardo, B., Shifrin, J. and Birmingham, W.P.: Name that tune: A pilot study in finding a melody from a sung query, *Journal of the American Society for Information Science and Technology*, Vol.55, No.4, pp.283–300 (2004).
9) Hu, N. and Dannenberg, R.B.: A Comparison of Melodic Database Retrieval Techniques Using Sung Queries, *Joint Conference on Digital Libraries*, pp.301–307 (2002).
10) Adams, N.H., et al.: Time Series Alignment for Music Information Retrieval, *Proc. ISMIR 2004* (2004).
11) 橋口博樹，西村拓一，張　建新，滝田順子，岡　隆一：モデル依存傾斜制限型の連続 DP を用いた鼻歌入力による楽曲信号のスポッティング検索，電子情報通信学会論文誌 D-II，Vol.J84-D-II, No.12, pp.2479–2488 (2001).
12) 酒向慎司，宮島千代美，徳田恵一，北村　正：隠れマルコフモデルに基づいた歌声合成システム，情報処理学会論文誌，Vol.45, No.3, pp.719–727 (2004).

13) Saitou, T., Unoki, M. and Akagi, M.: Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis, *Speech Communication*, Vol.46, pp.405–417 (2005).
14) de Cheveigné, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America*, Vol.111, No.4, pp.1917–1930 (2002).
15) 後藤真孝，橋口博樹，西村拓一，岡　隆一：RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol.45, No.3, pp.728–738 (2004).
16) Goto, M.: AIST Annotation for the RWC Music Database, *Proc. ISMIR 2006* (2006).
17) 後藤真孝，西村拓一：AIST ハミングデータベース：歌声研究用音楽データベース，情報処理学会音楽情報科学研究会研究報告，Vol.2005, No.82, pp.7–12 (2005).

(平成 20 年 2 月 21 日受付)
(平成 20 年 9 月 10 日採録)

大石　康智（学生会員）

2004 年名古屋大学工学部電気電子情報工学科卒業．2006 年同大学大学院情報科学研究科博士前期課程修了．現在，同大学院情報科学研究科博士後期課程在学中．音楽情報処理，音声言語情報処理に興味を持つ．2005 年日本音響学会ポスター賞受賞．日本音響学会学生会員．

後藤　真孝（正会員）

1998 年早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．同年電子技術総合研究所（2001 年に産業技術総合研究所に改組）に入所し，現在，主任研究員．2000 年から 2003 年まで科学技術振興事業団さきがけ研究 21 研究員，2005 年から筑波大学大学院准教授（連携大学院），2008 年から統計数理研究所客員准教授を兼任．音楽情報処理，音声言語情報処理等に興味を持つ．2001 年日本音響学会粟屋潔学術奨励賞，2005 年情報処理学会論文賞，2007 年第 6 回ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞，2008 年平成 20 年度科学技術分野の文部科学大臣表彰若手科学者賞等 22 件受賞．電子情報通信学会，日本音響学会，日本音楽知覚認知学会各会員．

伊藤　克亘（正会員）

博士（工学）．1993 年電子技術総合研究所入所．2003 年名古屋大学大学院情報科学研究科助教授．2006 年法政大学情報科学部教授．現在に至る．音声を主とした自然言語全般に興味を持つ．

武田　一哉（正会員）

1985 年名古屋大学大学院工学研究科修了，同年国際電信電話株式会社（現 KDDI）入社．1986 年国際電気通信基礎技術研究所（ATR）出向．1990 年 KDD 研究所復職（この間 1988 年から 1989 年まで米国 MIT 滞在研究員）．1995 年名古屋大学工学部助教授．2003 年名古屋大学情報科学研究科教授．この間，音声コーパス，音声合成，音声認識，音響信号処理，行動信号処理の研究・教育に従事．日本音響学会，電子情報通信学会，IEEE 各会員．