相平面を利用した歌声の F0 軌跡の新しい表現方法

A novel representation of sung melodic contour using phase plane

大石康智 [1]　後藤真孝 [2]　伊藤克亘 [3]　武田一哉 [1]
Yasunori Ohishi　Masataka Goto　Katunobu Itou　Kazuya Takeda

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

1　はじめに

我々は，歌声の旋律 (歌唱者の意図する音高列) と，ビブラートやオーバーシュートのような歌声特有の動的変動を特徴付ける信号モデルの構築を目指している．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一つであり，現在様々な研究がされているが，歌声の動的変動，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，観測される基本周波数 (F0) 軌跡から歌唱者の意図する音高列を正しく推定することが必要とされる．従来は，F0 軌跡を音高と音長を表すシンボル列に変換し，ngram モデルのような離散的な確率表現を利用して照合が行われたが，歌声が歌詞付きであったり，さらに動的変動を含むと，旋律を正しくシンボル列で表現することが難しい [1]．F0 軌跡そのものを DTW によって照合する方法が提案されているが，それでも歌声の動的変動の影響を受けて検索性能が低下してしまう [2]．また，F0 制御モデルを利用した自然性かつ明瞭性のある歌声合成が実現されているが，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 [3]．

そこで本報告では，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を特徴付けるための新しい表現方法を提案する．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も提案する．提案手法を利用したハミング検索実験を行ったところ，従来の F0 軌跡の DTW 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．

2　相平面における歌声の F0 軌跡

図 1(b) は，図 1(a) の歌声の F0 軌跡を相平面 $\vec{f}(y(t), \dot{y}(t))$ に図示した例である．ここで $y(t)$ は時刻 t の F0 を表し，de Cheveigne らの提案した YIN[4] を利用して 10ms ごとに推定する．なお，周波数単位 Hz を対数スケールの周波数単位 cent に変換する．一方，$\dot{y}(t)$ は時刻 t の F0 の時間微分を表し，微小区間 (50ms) の F0 の回帰係数 ΔF0 で近似する．以上で求めた $y(t)$ と $\dot{y}(t)$ を時刻 t の観測ベクトル $\mathbf{y}_t$ 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が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．これらのアトラクタの位置は，歌唱者が意図する音高目標値に対応する．一方，アトラクタにいたるまでの渦軌跡には，歌声の動的変動が現れる．例えば，ビブラートは，アトラクタを中心にその周りで楕円を描く軌跡として観測される．また，音高が遷移する時に目的音高より大きく振れてしまうオーバーシュートは，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．

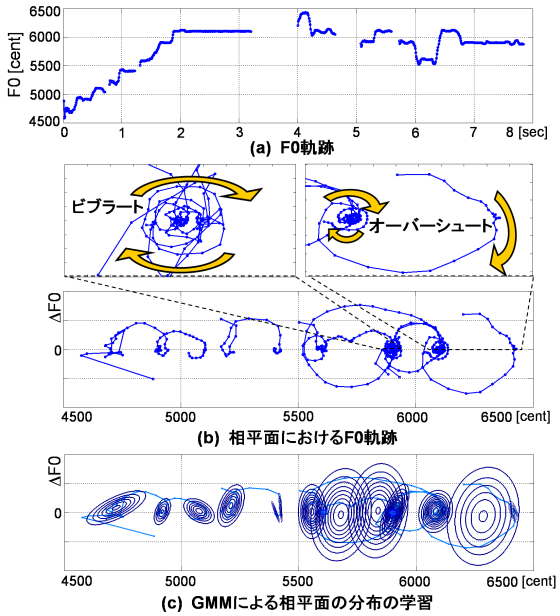

図 1　F0-ΔF0 の相平面に表現される歌声の F0 軌跡：音高遷移が，複数のアトラクタとそれらを遷移する動きによって表現される．ビブラート，オーバーシュートが，楕円または螺旋を描く軌跡によって表現される．(c) は相平面における F0 軌跡の分布を GMM によって学習した結果である．各ガウス分布の頂点がアトラクタの位置に対応する．

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 軌跡から歌唱者の意図する音高目標値の時系列を推定する．図 1(c) より GMM の各ガウス分布の頂点がアトラクタの位置に対応することから，図 2 に示すように，F0 軌跡をフレーム化処理し，各フレームごとに GMM を学習してアトラクタの位置，つまり音高目標値を推定することを考える．n 番目のフレームにおける観測ベクトルの集合を $\mathcal{Y}^{(n)}$，推定する GMM のパラメータを $\mathbf{\Theta}^{(n)}$ とする．GMM の混合数は M とする．このとき，$\mathcal{Y}^{(n)}$ が与えられたもとでの $\mathbf{\Theta}^{(n)}$ の事後確率 $p(\mathbf{\Theta}^{(n)}|\mathcal{Y}^{(n)})$ を最大にする $\hat{\mathbf{\Theta}}^{(n)}$ を，EM アルゴリズムによって求める．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 の観測ベクトル集合が従うアトラクタの位置 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

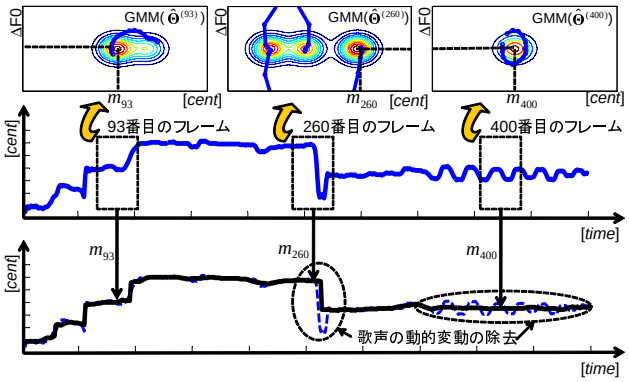

図 2　歌唱者の意図する音高目標値の推定手法：フレームごとに相平面の F0 軌跡を作成し，その分布を GMM で学習する．確率密度が最大となる F0 の値 m_n をフレーム n における音高目標値とする．結果的に歌声の動的変動が除去される．

が難しいので，相平面を格子に区切り，数値的に求める．図 2 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．ビブラートやオーバーシュートのような動的変動が除去され，歌唱者が本来歌おうとする音高列が推定される．この結果は，例えば，ハミング検索における旋律の距離尺度に利用できる．従来の F0 軌跡間の DTW では，動的変動の影響を受けて検索が難しい歌声に対して，類似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を推定し，その結果を利用してハミング検索実験を行う．ユーザによって入力される歌声 (入力信号) と楽曲データベースの旋律 (参照信号) 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，DTW によって類似距離を計算する．また従来法として，F0 軌跡間の DTW による検索性能も評価する．

「RWC 研究用音楽データベース：ポピュラー音楽」(RWCMDB-P-2001)[5] の計 100 曲から，歌唱の出だしの部分と盛り上がる主題の部分の 2 箇所を切り出し，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．本来ならばこれらの信号から F0 を推定することが望ましいが，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 [6] を用いた．

歌声研究用音楽データベース「AIST ハミングデータベース」[7] の一部である，日本人歌唱者 75 名 (男性 37 名，女性 38 名) が，上記の 200 種類の参照信号のうち 50 種類を歌詞付き，伴奏なし，自由なテンポで歌唱した歌声を入力信号として利用する．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，38 種類の旋律を歌唱した 1,350 サンプルを実験で利用する．

提案手法のフレーム長は 200ms，フレームシフト長は 10ms とした．各フレームごとに推定する GMM の混合数は $M = 2$ と固定した．

5　実験結果

再現率と適合率の調和平均で計算される F 値で検索性能を評価する．図 3 は，DTW による類似距離が閾値 70 以下であるものを検索結果としたときの F 値を，入力信号の旋律ごとに計算した結果である．「P005 出だし」は，楽曲番号 P005 の出だしの旋律を歌唱した入力信号の集合を意味する．旋律によって F 値の改善度合は異な

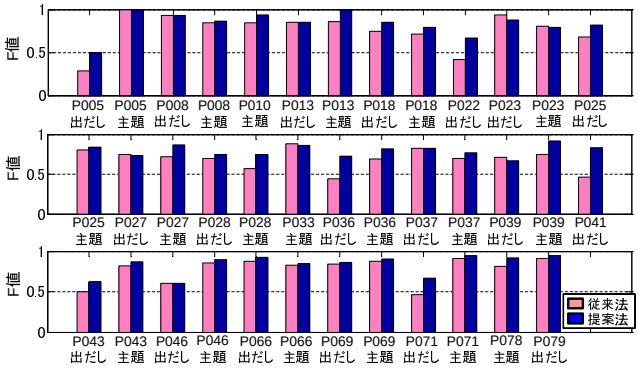

図 3　F 値による検索性能の評価

るが，全体を平均すると 0.632 から 0.711 に改善された．類似距離の閾値を 70 以下に設定しても F 値の改善が確認された．したがって，提案手法によって入力信号と参照信号の DTW の局所類似距離が改善され，適切に検索結果を絞り込むことが可能であることがわかった．

ただ，歌唱する旋律によっては F 値が低下してしまう原因の一つとして，実験条件のフレーム長や GMM の混合数の固定が影響していると考えられる．音符を多く含む速いパッセージの旋律もあれば，ゆるやかな旋律もあるため，現在の実験条件では場合によってはフレーム中の音符を削除してしまうことが考えられる．フレームに含まれる音符の数を推定して適切な混合数で GMM を学習すること，また動的にフレーム長を変動させながら GMM を学習するなどの発展が考えられる．

6　まとめ

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歌声の動的変動を明確に可視化することができる．そこで，相平面の F0 軌跡から音高目標値を推定する手法を提案した．推定された音高目標値の時系列を利用することによって，ハミング検索の性能改善を確認した．今後の課題は，先に述べた提案手法の実験条件となるパラメータを適切に決定することである．また，動的変動を表現する制御モデルを導入することによって，音高目標値と動的変動の制御パラメータを同時に推定することを考えれば，音高目標値の推定精度もさらに向上すると考えられる．この点は現在検討している段階にあり，実現されればハミング検索だけでなく，制御パラメータを利用した歌唱スタイルに基づく楽曲検索や歌声合成への応用が期待される．

参考文献

[1] Dannenberg, R. et al.: A Comparative Evaluation of Search Techniques for Query-by-Humming Using the MUSART Testbed, *JASIST*, Vol. 58, No. 5, pp. 687–701 (2007).

[2] Hu, N. et al.: A Comparison of Melodic Database Retrieval Techniques Using Sung Queries, *Joint Conference on Digital Libraries*, pp. 301–307 (2002).

[3] Saitou, T. et al.: Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis, *Speech Communication*, Vol. 46, pp. 405–417 (2005).

[4] de Cheveigne, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music, *JASA*, Vol. 111, No. 4, pp. 1917–1930 (2002).

[5] 後藤真孝ほか：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).

[6] Goto, M.: AIST Annotation for the RWC Music Database, *ISMIR 2006* (2006).

[7] 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).