確率的定式化と歌声の統計的モデル化による ボーカルパートの基本周波数推定手法

An F0 Estimation Method of Vocal Part
based on Stochastic Formulation and Statistical Modeling of Singing Voice

藤原 弘将 [1] 後藤 真孝 [1] 奥乃 博 [2]
Hiromasa Fujihara Masataka Goto Hiroshi G. Okuno

産業技術総合研究所 [1]
National Institute of Advanced Industrial Science and Technology (AIST)
京都大学 [2]
Kyoto University

1 まえがき

歌声はポピュラー音楽をはじめとして多くのジャンルの音楽で重要な役割を果たしている．そのため，ボーカルパートの基本周波数 (F0) 推定は，計算機で音楽を認識する上で重要な課題である．本稿では，市販 CD 等の歌声と伴奏音を同時に含む楽曲を対象とし，ボーカルパートの F0 を自動推定する手法の概要について述べる[1]．本研究では，複数の歌手が交互にボーカルパートを歌う楽曲やメインのボーカルパートと同時にコーラスなどのパートが歌われる楽曲も対象に含める．一方，メインのボーカルパートが同時に複数の歌手によって，異なる音高で歌われることは無いと仮定する．この技術は，ボーカルパートの自動採譜やカラオケトラックの自動作成などに応用することができる．従来のボーカルパートの F0 推定手法 [2, 3] はボーカルパートを選び出す際に，F0 の優勢さや基本周波数の変化の仕方のみを手がかりとして用い，音響的特徴は考慮していなかった．

本研究では，歌声の音響的特徴を混合ガウス分布（GMM）でモデル化することで，対象パートをボーカルパートに限定する．これにより，高精度なボーカルパート F0 推定を実現する．まず確率的定式化により，ボーカルパート F0 推定の問題を多重 F0 解析の問題と音源認識の問題に分割する．多重 F0 解析の問題とは，複数の高調波構造が混合したスペクトルから，混合前のそれぞれの高調波構造の F0 を推定する問題である．音源認識の問題とは，スペクトル中のある F0 の音源（ここでは歌声かどうか）を推定する問題であり，本研究では，歌声と非歌声をモデル化した GMM により実現する．

2 ボーカルパート F0 推定問題の確率的定式化

各時刻 t ($t = 1, \cdots, T$) における F0，スペクトル，音源の種類を確率変数としてそれぞれ，$F = \{f_t | t = 1, \cdots, T\}$，$\Psi = \{\psi_t | t = 1, \cdots, T\}$，$\Lambda = \{\lambda_t | t = 1, \cdots, T\}$ と定義する．また，音源の種類として歌声と歌声以外の 2 種類を考える．

ボーカルパートの F0 推定問題は，スペクトルを観測し，全ての時刻で音源の種類が歌声であるとした場合に，次式を最大化する F0 系列，$\hat{F}$，を求めることである．

$$\hat{F} = \underset{F}{\operatorname{argmax}} \log p(F|\Psi, \Lambda) \qquad (1)$$

[1]本稿で述べる技術の詳細や，より細かい実験結果は文献 [1] に書かれている．

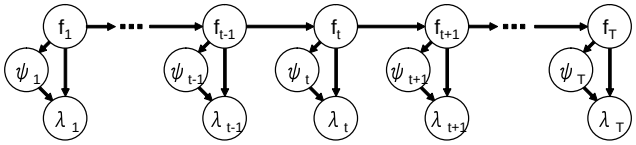

図 1 F，Ψ，Λ の確率的依存関係

ただし，S_V は，全ての時刻で音源が歌声であることを表す．ここで，確率変数 F，Ψ，Λ の確率的依存関係が，図 1 式 (1) のように表現できると仮定すると，式 (1) は，

$$\hat{F} = \underset{F}{\operatorname{argmax}} \prod_{t=1}^{T} p(\lambda_t|f_t, \psi_t) p(\psi_t|f_t) p(f_t|f_{t-1}) \qquad (2)$$

のように展開することができる．

右辺第 1 項 $p(\lambda_t|f_t, \psi_t)$ は，スペクトル中のある F0 の音の音源が歌声である確率を意味し，歌声・非歌声確率と呼ぶ．これは，音源認識の問題と捉えることができる．右辺第 2 項 $p(\psi_t|f_t)$ は，スペクトル中にある F0 の音が存在するかどうかを表し，F0 尤度と呼ぶ．これは，多重 F0 解析の問題と捉えることができる．右辺第 3 項 $p(f_t|f_{t-1})$ は，F0 軌跡の変化に関する制約を表現し，F0 遷移確率と呼ぶ．これらの 3 つの条件付き確率を適切に定めることで，$\hat{F}$ は Viterbi アルゴリズムを用いて効率的に推定することができる．このように，ボーカルパートの F0 推定問題を，音源認識の問題と多重 F0 推定の問題に分割して考えることで，F0 推定の際にボーカルパートのみに着目することを可能にした．

3 確率計算

2 節で述べた定式化によって F0 を推定するためには歌声・非歌声確率 $p(\lambda_i|f_i, \psi_i)$，F0 尤度 $p(\psi_i|f_i)$，F0 遷移確率 $p(f_i|f_{i-1})$ を適切に定義する必要がある．本節ではこれらの確率関数の計算方法について述べる．

歌声・非歌声確率 歌声・非歌声確率 $p(\lambda_i|f_i, \psi_i)$ は，観測スペクトル中で特定の F0 の音が歌声であるかどうかを表現する．これは，音源が歌声か歌声でないかを推定するという意味で，音源認識の問題と捉えることができる．本研究では,，歌声と歌声以外の音 (非歌声) をモデル化した 2 つの GMM の尤度を用いて歌声確率を計算する．図 2 に本手法の概要を示す．観測スペクトルに対して，各 F0 それぞれの高調波構造を分離し，正弦波重畳モデルを用いて再合成する．これにより，各 F0 ごと

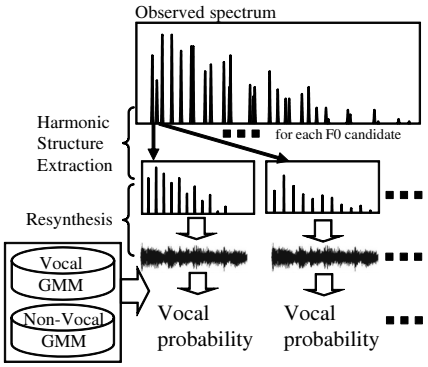

図 2　歌声確率の計算

にそれぞれ分離信号が得られる．さらに，これらの分離信号から特徴量を抽出し，GMM を用いて歌声・非歌声確率を計算する．GMM で使用する特徴量として，線形予測メルケプストラム係数 (LPMCC) と F0 の微分係数（ΔF0）を使用する．

F0 尤度　F0 尤度 $p(\psi_i|f_i)$ は，観測スペクトル ψ_i 中に f_i を基本周波数とする高調波構造が存在する可能性がどの程度あるかを表す．F0 尤度の計算には，後藤の PreFEst[4] を用いる．PreFEst は多重奏の音響信号中の最も優勢な高調波構造を持つ F0 を推定する手法である．具体的には，各時刻の F0 候補を推定する front-end, core と最も優勢な F0 軌跡を追跡する back-end からなる．本研究では front-end と core により推定された各時刻の F0 候補とその重みを F0 尤度として用いる．

F0 遷移確率　F0 遷移確率 $p(f_i|f_{i-1})$ とは，F0 の時間的連続性に関する制約を表す．本研究では，ラプラス分布を用いて

$$p(f_i|f_{i-1}) = \frac{1}{2b}\exp\left(-\frac{|f_i - f_{i-1}|}{b}\right) \qquad (3)$$

のように定義する．ただし，f_i，f_{i-1} は cent の単位で表される周波数である[2]．b はラプラス分布のスケールを規定するパラメータで，本研究では，$p(f_i|f_{i-1})$ の標準偏差が 150 cent になるように設定した．

4　評価実験

本手法の有効性を確認するため，音源を考慮しない F0 推定手法である PreFEst と比較し，評価実験を行った．実験には,「RWC 研究用音楽データベース: ポピュラー音楽 (RWC-MDB-P-2001)」中の全曲（100 曲）を用いた．評価は，10 fold cross-validation 法を用いて行われた．正解の判断基準として，正解のメロディの音高を人間が手作業でアノテーションしたデータを用いた．正解率として，歌声が存在する区間のみを用い，楽曲の全体長に対する正解区間長の割合を計算した．正しいと判定する周波数差の基準は，50cent 以下とした．歌声 GMM のパラメータは，上記のメロディのアノテーションデータにおいてメロディの音高がアノテーションされている区間を用いて，正解の F0 を用いて分離した音響信号より計算された特徴量を用いて推定した．非歌声 GMM のパラメータは，メロディのアノテーションデータにおいてメロディの音高がアノテーションされていない区間を用いて，ミックスダウンされたデータから PreFEst を用いて推定された最も優勢な F0 系列を用いて分離した音響信号から計算された特徴量を用いて推定した．

PreFEst の平均正解率が 76.2%なのに対し，提案手法の平均正解率が 81.1%であった．つまり，本手法を用いることで正解率が 4.9 ポイント向上し，誤り率を 20.5%削減できた．これにより，ボーカルパートに特化することの効果が確認できた．

本手法と PreFEst の推定結果を比較すると，PreFEst では歌声が徐々に小さくなる箇所で歌声の F0 を追跡しきれずに途中で他の楽器の F0 を追跡してしまう場合があったが，本手法ではそのような場面でも歌声の F0 を正しく追跡できている場面が散見された．また，PreFEst を用いて F0 候補を推定した段階で，低域に歌声とは無関係な F0 候補が多く見られた．PreFEst では，音源を仮定していないため，歌声が存在する区間でも低域のノイズの F0 を追跡してしまうことが多かったが，本手法ではそのような低域のノイズの F0 は歌声確率が低くなるため，そのようなノイズに惑わされることなく歌声の F0 を正しく追跡できる場合が多かった．

5　まとめ

本論文では，多重奏の音響信号から，ボーカルパートの F0 を推定する手法について述べた．本手法では，ボーカルパートの F0 推定の問題を確率的に定式化し多重 F0 解析と音源認識の問題に分割することで，推定結果をボーカルパートに限定することを可能にした．さらに，確率的定式化によりボーカルパートの F0 推定の問題を各確率関数の設計の問題に帰着させたことで，今後の手法の改良の見通しが立てやすいという利点がある．

現在は歌声区間の推定は行っていないため，間奏区間でも何らかの F0 を結果として出力しているが，今後は歌声区間推定を統合し F0 推定と同時に間奏区間を推定することが課題となる．さらに，本手法の枠組みは歌声以外の楽器にも容易に拡張できるものとなっているため，今後はこの枠組みの中で歌声以外の特定楽器パートの F0 推定に応用していく予定である．本研究の一部は，科研費，CREST の支援を受けた．

[2] cent は本来ある値を基準とした相対音高を表す単位であるが，ここでは 2 つの周波数の差として用いるのみなので，基準となる値は任意でかまわない．

参考文献

[1] 藤原 弘将 他, 歌声 GMM とビタビ探索を用いた多重奏中のボーカルパートに限定した基本周波数推定手法, 情処音情研, Vol.2007, No.81, pp.119-126 (2007).

[2] Ryyänanen *et al.*, Transcription of the Singing Melody in Polyphonic Music, *ISMIR*, pp.222–227 (2006).

[3] Li *et al.*, Detection Pitch of Singing Voice in Polyphonic Audio, in *Proc. ICASSP*, pp.17–20 (2005).

[4] Goto, A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detection Melody and Bass Lines in Real-World Audio Signals, *Spe. Comm.*, vol. 43, no. 4, pp. 311–329 (2004).