2007－MUS－71 (20)
2007／8／2

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

藤原 弘将† 後藤 真孝† 奥乃 博‡

† 産業技術総合研究所 ‡ 京都大学大学院 情報学研究科 知能情報学専攻

本稿では，混合音中のボーカルパートの基本周波数 (F0) を推定する手法について述べる．ボーカルパートは多くのジャンルの音楽で主要な役割を果たしており，ボーカルパートのF0推定は様々な用途に応用出来る．我々は，確率的定式化により，ボーカルパートのF0推定の問題を多重ピッチ解析問題と音源 (歌声かどうか) 認識問題に帰着させる．音源認識問題は，歌声/非歌声を表現する混合ガウス分布 (GMM) を用いて，歌声確率を計算することで実現する．最後に，これらの確率的問題を最大化するF0の系列をビタビアルゴリズムによって推定する．評価実験により，F0推定精度が75.4%から78.3%に向上し，誤り率を13.8%削減することを確認した．

An F0 estimation method for detecting vocal part in polyphonic music by using vocal GMM and Viterbi search

HIROMASA FUJIHARA†, MASATAKA GOTO† and HIROSHI G. OKUNO‡

† National Institute of Advanced Industrial Science and Technology (AIST)
‡ Dept. of Intelligence Science and Technology, Graduate School of Infomatics, Kyoto University

This paper describes a method for estimating F0s of vocal part from polyphonic audio signals. Because melody is sung by a singer in many musical pieces, the estimation of F0s of the vocal part is useful for many applications. We separate the problem of estimating F0s of vocal into multiple-F0 estimation problem and sound source (vocal or not) recognition problem. To deal with the sound source recognition problem, we evaluate the vocal probability by using vocal and non-vocal Gaussian mixture models (GMMs), Finally, we estimate an F0 trajectory that maximize these stochastic problems. based on Viterbi search. Experimental results show that our method improves estimation accuracy from 75.4% to 78.3%, which is 13.8% reduction of misestimation.

1. はじめに

ポピュラー音楽を始めとする多くのジャンルの音楽では，ボーカルの歌う歌声は中心的な役割を果たしている．歌声を計算機で自動的に理解することが出来れば，音楽情報検索システムをはじめとして，様々な用途で有用である．しかし，通常歌声はその他の伴奏音と混ざった状態で提供されるため，その理解は計算機には困難であった．我々はこれまで，多重奏中の歌声の理解を目的とし，声質を理解するための歌手名の同定[1]，歌詞を理解するための音楽と歌詞の時間的対応付け[2]の研究を行ってきた．そして，次の段階として歌詞と並んで楽曲を構成する最も重要な要素の一つである旋律に着目し，多重奏中のボーカルパートの基本周波数 (F0) 推定について研究した[3]．本稿では，ボーカルパートのF0推定手法について改めて述べると共に，より詳細に評価実験を行った結果について報告する．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．

従来から，多重奏中のF0推定に関する研究は多くあった[4)-9)]．多重奏の音楽からF0推定する場合に，どのパートのF0を推定するかという問題が発生する．従来のアプローチは，全ての音のF0推定を目指すものと，ある特定のF0推定を目指すものの2種類があった．全ての音のF0推定を目指す研究では，音源分離ドラムなどの非調波音が混在していると精度が大きく低下するなどの問題があった．特定のF0推定を目指した研究では，優勢さ，音色の類似性，拍子，F0の連続性などを手がかりに，推定対象のパートを選択し，パートの音源は考慮しているものは少数であった．そのため，歌声と同時に演奏される楽器のF0を歌声と誤って推定してしまうことがあった[4]．

Tyynanenら[8]は，多重奏中の歌声のF0推定の問題

に取り組んでいた．彼らは低レベルの音響的特徴と，高レベルの音楽的文脈の情報を組み合わせて，歌声のパートを追跡していた．しかし，音響的特徴として用いたものが，基本周波数の変化の仕方や強度の情報のみで歌声の手がかりとしては不十分であった．Li ら[7]は，自己相関に基づく方法で，多重奏中の歌声の F0 推定の問題に取り組んでいたが，歌声かどうかの判別は行わず，高域で最も優勢なピークを選択していた．糸山ら[10]は，楽器音を対象に特定パートの自動採譜の研究に取り組んでいたが，彼らは，全ての単音についての F0 推定と単音形成を予め行った後に，それらの単音中から特定の音源のものを選別するというアプローチであった．しかし，このアプローチでは推定対象の音源以外の単音形成にも高い推定精度が要求されるという問題があった．また，楽曲に存在する全ての楽器が既知でなければならないという制約があった．

本研究では，対象とするパートをボーカルパートに限定することで，より高精度な F0 推定の実現を目指す．我々は，まず確率的定式化により，特定パート F0 推定の問題を多重 F0 解析問題と音源認識の問題に分割して扱うことを可能にする．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スペクトルの F0 を推定する問題であり，従来手法を用いる．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混合ガウス分布 (GMM) により実現する．最後に，ビタビアルゴリズムを用いて効率的な方法で，この確率的定式化の解となる F0 軌跡を推定する．

2. ボーカルパートの F0 推定手法

本論文では，与えられた音楽音響信号中のボーカルパートの F0 を推定する問題を扱う．対象とするデータは，市販 CD 等の歌声と伴奏音を同時に含む楽曲である．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する．

本研究では，ボーカルパート F0 推定の問題を確率を用いて定式化する．これにより，ボーカルパート F0 推定の問題を，**F0 尤度**，**歌声確率**，**F0 遷移確率**の 3 つの確率の設計の問題に帰着させることが出来る．F0 尤度の計算は，多重 F0 解析の問題であり，従来手法を用いて計算する．歌声確率の計算とは，スペクトル中のある F0 の音源が歌声であるかどうかを判定する問題であり，歌声，非歌声を表現する GMM を用いて計算する．F0 遷移確率は，F0 がなめらかに変化するための制約であり，ガウス関数を用いて設計する．このように，多重 F0 解析の問題と音源認識の問題に分割して考えることで，多重奏中の特定パートの F0 を推定することを可能にした．

2.1 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_t，ψ_t，λ_t と定義する．さらに，

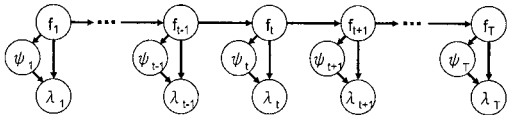


図 1 F，Ψ，Λ の確率的依存関係

$$F=\{f_t|t=1,\cdots,T\}, \tag{1}$$
$$\Psi=\{\psi_t|t=1,\cdots,T\}, \tag{2}$$
$$\Lambda=\{\lambda_t|t=1,\cdots,T\}, \tag{3}$$

と定義する．ここでは，音源の種類として歌声 (s_V) と歌声以外 (s_N) の 2 種類を考える．つまり，$\lambda_t\in\{s_V,s_N\}$ である．

特定パートの F0 推定問題は，スペクトルの時系列 $O=\{o_t|t=1,\cdots,T\}$ を観測し，全ての時刻で音源の種類が歌声 (s_V) 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．

$$\hat{F}=\underset{F}{\operatorname{argmax}}\log p(F|\Psi=O,\Lambda=S_V) \tag{4}$$

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 のように表現できると仮定すると，式 4 は，

$$\begin{aligned}\hat{F}&=\underset{F}{\operatorname{argmax}}\{\log p(\Lambda=S_V|F,\Psi=O)\\&\quad+\log p(\Psi=O|F)+\log p(F)\}\end{aligned} \tag{5}$$

$$\begin{aligned}&=\underset{F}{\operatorname{argmax}}\left\{\sum_{t=1}^{T}\log p(\lambda_t=s_V|f_t,\psi_t=o_t)\right.\\&\quad\left.+\sum_{t=1}^{T}\log p(\psi_t=o_t|f_t)+\sum_{t=1}^{T}\log p(f_t|f_{t-1})\right\}\end{aligned} \tag{6}$$

と分解することが出来る．

右辺第 1 項 $p(\lambda_i|f_i,\psi_i)$，は，スペクトル中にのある F0 を基本周波数とする音が存在した場合に，その音の音源が歌声 (または歌声以外の音) 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 $p(\psi_i|f_i)$，は，スペクトル中にある F0 を基本周波数とする音が存在するかどうかを表す尤度を意味し，**F0 尤度**と呼ぶ．これは，多重 F0 解析の問題と捉えることが出来る．右辺第 3 項 $p(f_i|f_{i-1})$，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．このようにして，特定パートの 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，これらの 3 つの条件付き確率を適切に定めることで，F0 推定の際に特定の音源に着目することを可能にした．これらの条件付き確率の計算方法は，3 節で述べる．

2.2 ビタビ探索による F0 軌跡の推定

式 (6) を最大化する F0 系列を求める．これは，Viterbi 探索に基づくアルゴリズムで，効率良く計算できる．まず，式 (6) に対して，確率・尤度の計算方式の違いによるスケールの違いを正規化するため，F0 尤度，歌声確率，F0 遷移確率の間に結合重みを導入する．

$$\hat{F} = \underset{F_T}{\operatorname{argmax}} \left\{ \alpha \sum_{t=1}^{T} \log p(\lambda_t = s_V | f_t, \psi_t = o_t) + \beta \sum_{t=1}^{T} \log p(\psi_t = o_t | f_t) + \sum_{t=1}^{T} \log p(f_t | f_{t-1}) \right\} \quad (7)$$

これは，音声認識での言語モデルと音響モデル間への結合重みの導入と同様の考え方である．本研究では，$\alpha = 0.3$，$\beta = 0.7$ と設定し，式 (6) のかわりに式 (7) を用いた．

式 (7) を直接計算することは困難であるため，以下の式に従って再帰的に計算する．まず，バックポインタ $B(t,f)$ と累積確率 $A(t,f)$ を導入する．バックポインタ $B(t,f)$ は，時刻 t に基本周波数 f であった場合の，時刻 $t-1$ での基本周波数の値を表す．累積確率 $A(t,f)$ は，時刻 t に基本周波数 f である確率である．

(1) 初期化

$$\forall f \; A(1,f) = \alpha \log p(\lambda_1 | f, \psi_1) + \beta \log p(\psi_1 | f) \quad (8)$$

(2) 再帰的計算 $(t = 2, \cdots, T)$

$$A(t,f) = \max_{f'} \{ A(t-1,f') + \alpha \log p(\lambda_t | f, \psi_t) + \beta \log p(\psi_t | f) + \log p(f | f') \} \quad (9)$$

$$B(t,f) = \underset{f'}{\operatorname{argmax}} \{ A(t-1,f') + \alpha \log p(\lambda_t | f, \psi_t) + \beta \log p(\psi_t | f) + \log p(f | f') \} \quad (10)$$

(3) バックトラック

以上で，全て時刻 t の基本周波数 f に対してバックポインタ $B(t,f)$ が計算された．最後に，$B(t,f)$ を後ろ向きに辿っていくことで，式 (7) を最大化する F0 の系列 $(\hat{F} = \{\hat{f}_1, \cdots, \hat{f}_T\})$ を得ることができる．

$$\hat{f}_T = \underset{f}{\operatorname{argmax}} A(T,f) \quad (11)$$

$$\hat{f}_t = B(\hat{f}_{t+1}) \quad (t = T-1, \cdots, 1) \quad (12)$$

2.3 リアルタイム処理

2.1 節の定式化では，全ての時刻のスペクトルが既知であるという条件で最尤な F0 軌跡を推定している．しかしこれでは，楽曲の最後まで入力しないとボーカルパートの F0 を推定することが出来ないので，リアルタイム処理が行えないという問題点がある．そこで，リアルタイム処理が必要な場合は式 (4) を改変して，

$$\begin{aligned} \hat{f}_t &= \underset{f_t}{\operatorname{argmax}} \log p(f_t | \psi_0 = o_0, \cdots, \psi_{t+N} = o_{t+N}, \\ &\quad \lambda_0 = s_V, \cdots, \lambda_{\tau+N} = s_V) \\ &= \underset{f_t}{\operatorname{argmax}} \log \int \cdots \int p(f_t \cdots f_{t+N} | \psi_t = o_t, \cdots, \\ &\quad \psi_{t+N} = o_{t+N}, \lambda_t = s_V, \cdots, \lambda_{t+N} = s_V) \\ &\quad df_{t+1} \cdots df_{t+N} \end{aligned} \quad (13)$$

とする．すなわち，F0 を求めたい時刻 t から N フレーム先までの情報のみを手がかりに，時刻 t の F0 を決定する．さらに，式 (13) を厳密に計算するためには時刻 t 以外の時刻で周辺化した確率を計算する必要があるが，本研究では計算の簡略化のためと，リアルタイム処理を行わない場合との共通性を確保するため，

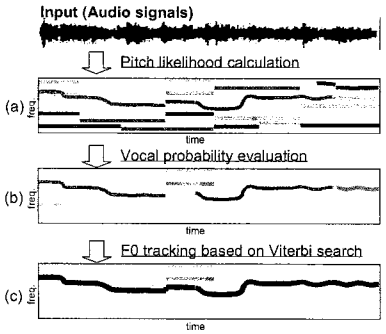


図 2　処理の流れ．

$$\hat{f}_t = \underset{f_t}{\operatorname{argmax}} \log p(f_t \cdots f_{t+N} | \psi_t = o_t, \cdots, \psi_{t+N} = o_{t+N}, \lambda_t = s_V, \cdots, \lambda_{t+N} = s_V) \quad (14)$$

のように近似する．式 (14) は，2.1 節，2.2 節と同様の処理で高速に計算することが出来る．

2.4 従来手法との関係

本節では，従来の混合音からの F0 推定と比較して，本稿の定式化がどのように位置づけられるかについて述べる．多重奏の音響信号からの F0 推定についての先行研究[4)-6),8)] の多くは，2 つの処理から成り立っていた．すなわち，多重ピッチ解析の技術を用いて複数の音が混ざったスペクトルから F0 の候補を推定する処理と，得られた候補の中から F0 の軌跡を，優勢さ，音色，拍子，F0 の連続性などの手がかりを用いて推定する処理である．

それに対し，本手法では，各時刻のスペクトルから得られた F0 の候補に対し，歌声確率を用いて歌声以外の音源の F0 に低い重みを付けていると解釈することが出来る．このようにして音源の種類を限定することで，より高精度に F0 を推定する．そして，各時刻の F0 候補から F0 軌跡を追跡する処理の 1 つの実現方法として，2.2 節で述べたビタビ探索を導入している．図 2 は，このような見方で考えた本手法の処理の流れである．

3. 確率計算

本節では，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 $p(f_i | f_{i-1})$ の具体的な計算方法について述べる．

3.1 歌声/非歌声確率

2.1 節で導入された歌声/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出来る．従来の歌声/非歌声推定手法[11)-13)] はパワーやゼロ交差，MFCC 等の特徴量を多重奏の音響信号から直接計算していた．そのため，混合音中の特定の F0 に着目して音源を推定することが出来なかった．

図 3 に本手法の概要を示す．本研究では，観測スペクトル中の全ての F0 について，高調波構造を分離し，正

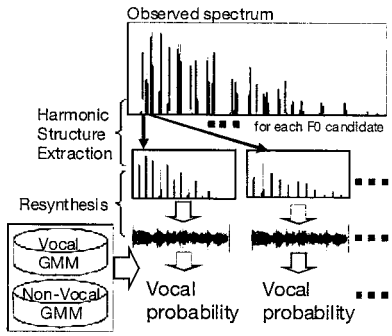


図 3 歌声確率の計算

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．

3.1.1 高調波構造の分離

メロディの高調波構造の各倍音成分のパワーを抽出する．各周波数成分の抽出には，前後 20 cent ずつの誤差を許容し，この範囲で最もパワーの大きなピークを抽出する．l 次倍音 ($l=1,\ldots,L$) のパワー A_l と周波数 F_l は，以下のように表される．

$$F_l = \operatorname*{argmax}_F |S(F)| \quad (l\overline{F_0}\cdot(1-2^{\frac{r}{1200}}) \le F \le l\overline{F_0}\cdot(1+2^{\frac{r}{1200}})), \quad (15)$$

$$A_l = |S(F_l)|, \quad (16)$$

ここで，$S(F)$ はスペクトルを，$\overline{F_0}$ は 抽出したい F0 を表す．

抽出された高調波構造を正弦波重畳モデル[14] に基づき再合成することで，分離音響信号を得る．再合成された音響信号 $s(t)$ は，

$$s(t) = \sum_{k=1}^{K} A_k \cos(\omega_k t), \quad (17)$$

と表わされる．ここで，A_k, ω_k はそれぞれ，k 次倍音のパワー，周波数を表わし，t は時間を表わす．

3.1.2 特徴抽出

歌声信頼度計算のための GMM で使用する特徴量は，歌声の音響的特徴を良く表現し，歌声以外の楽器と歌声との差を際だたせるものが望ましい．本研究では，歌声の音響的特徴を表現する特徴量として線形予測メルケプストラム係数 (LPMCC) を，歌声の動的な特性を表現する特徴量として F0 の微分係数 (ΔF0) を導入し，これらを並べて 1 つの特徴ベクトルとしたものを用いる．

- LPC メルケプストラム (LPMCC)[15]:
  LPMCC は，LPC スペクトルから計算されたメルケプストラム係数で，歌声の音響的特徴を表現する．音声や音楽から抽出する音響的特徴量として，メル周波数ケプストラム係数 (MFCC)[16),17] がよく用いられてきた．本研究では，LPC スペクトルから MFCC を計算することで LPMCC を抽出した．
- ΔF0:
  歌声の動的な性質を表現する特徴量として，F0 の微分係数 (ΔF0)[18] を用いた．歌声は他の楽器音と比較して，ビブラートなどに起因する時間変動が多いので，F0 の軌跡の傾きを表す ΔF0 は，歌声と非歌声の識別に適していると考えられる．
  ΔF0 の計算には，次式のように 5 フレーム間の回帰係数を用いた．

$$\Delta f[t] = \frac{\sum_{k=-2}^{2} k\cdot f[t+k]}{\sum_{k=-2}^{2} k^2}, \quad (18)$$

ここで，$f[t]$ は，時刻 t における周波数 (単位: cent) であるとする．
式 (18) を計算するためには，ある時刻 t のそれぞれの F0 f について，前後の時刻 $t, t-1$ において接続する F0 が必要である．これは，F0 尤度 $p(\psi|f)$ を用いて以下のように計算する．ある時刻 t の F0 f_t に隣接する時刻 $t+1$(または $t-1$) における F0 の値 f_{t+1}（または f_{t-1}）は，

$$f_{t+1} = \operatorname*{argmax}_{f_{t+1}} p(\psi_{t+1}|f_{t+1})\mathcal{N}(f_t; f_{t+1}, 100) \quad (19)$$

$$f_{t-1} = \operatorname*{argmax}_{f_{t-1}} p(\psi_{t-1}|f_{t-1})\mathcal{N}(f_t; f_{t-1}, 100) \quad (20)$$

である．ここで，$\mathcal{N}(x; m, \sigma^2)$ は平均 x，標準偏差 σ のガウス分布を表す．

3.1.3 確率計算

歌声・非歌声確率の計算では．歌声が存在する区間から抽出された特徴量で学習した歌声 GMM θ_V と，伴奏区間から抽出された特徴量で学習した非歌声 GMM θ_N を用いる．すなわち，音源の状態 (歌声 s_V または歌声以外 s_N) を観測した際のスペクトル ψ_t と 基本周波数 f の同時確率 $p(\psi_t, f|s_V)$，$p(\psi_t, f|s_N)$ を，歌声 GMM の尤度 $\mathcal{N}_{GMM}(x;\theta_V)$ と非歌声 GMM の尤度 $\mathcal{N}_{GMM}(x;\theta_N)$ を用いて，

$$p(\psi_t, f|s_V) = \mathcal{N}_{GMM}(x(\psi_t, f); \theta_V) \quad (21)$$

$$p(\psi_t, f|s_N) = \mathcal{N}_{GMM}(x(\psi_t, f); \theta_N) \quad (22)$$

と定義する．これらを用いて歌声/非歌声確率は，

$$p(s_V|\psi_t, f) = \frac{p(\psi_t, f|s_V)}{p(\psi_t, f|s_V) + p(\psi_t, f|s_N)} \quad (23)$$

$$p(s_N|\psi_t, f) = \frac{p(\psi_t, f|s_N)}{p(\psi_t, f|s_V) + p(\psi_t, f|s_N)} \quad (24)$$

で表される．ただし，$x(\psi_t, f)$ は時刻 t のスペクトル ψ_t の基本周波数 f の周波数成分を伴奏音抑制によって分離した信号から計算された特徴量を表す．

3.2 F0 尤度

F0 尤度の計算には，後藤の PreFEst[4] を用いる．PreFEst は，front-end，core，back-end の 3 つの処理からなるが，本研究では各時刻の周波数成分の候補を求める front-end と core のみを用いる．back-end は，core によって得られた各時刻の周波数成分の候補の中から，最も優勢な F0 軌跡を追跡する処理であり，本手法では用いない．

以下に，PreFEst-core の概要を記す．パワースペクトル $\psi(f)$ が与えられた時，以後の確率的処理を可能に

表 1 歌声/非歌声 GMM の学習データ

Name	Gender	Piece Number
Shingo Katsuta	M	027
Yoshinori Hatae	M	037
Masaki Kuehara	M	032, 078
Hiroshi Sekiya	M	049, 051
Katsuyuki Ozawa	M	015, 041
Masashi Hashimoto	M	056, 057
Satoshi Kumasaka	M	047
Konbu	F	013
Eri Ichikawa	F	020
Tomoko Nitta	F	026
Kaburagi Akiko	F	055
Yuzu Iijima	F	060
Reiko Sato	F	063
Donna Burke	F	081, 091, 093, 097

するため，周波数成分を確率密度関数 (PDF) として，

$$p_\psi(f) = \frac{\psi(f)}{\int_{-\infty}^{\infty}\psi(f)df} \tag{25}$$

のように表現する．そして，観測された PDF が次式で表されるように音モデルの重み付き混合から生成されたと考える．

$$p(f|\theta) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w(F)p(f|F)dF, \tag{26}$$

$$\theta = \{w(f)|\mathrm{Fl} \le f \le \mathrm{Fh}\} \tag{27}$$

ここで，$p(f|F)$ は各 F0 の音モデルの PDF であり，Fh と Fl は考慮する周波数範囲の下限と上限を表す．また，$w(f)$ は音モデルの重みで，

$$\int_{\mathrm{Fh}_i}^{\mathrm{Fl}_i} w(f)df = 1 \tag{28}$$

を満たす．音モデルは典型的な高調波構造を表現した確率分布である．そして，EM アルゴリズムを用いて $w(f)$ を推定し，それを F0 の確率密度関数と解釈する．本研究では，この F0 の確率密度関数を F0 尤度関数として用いる．すなわち，

$$p(\psi|f) = w(f) \tag{29}$$

である．

3.3 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し，

$$p(f_i|f_{i-1}) = \mathcal{N}(f_i; f_{i-1}, W_V), \tag{30}$$

のように定義する．ここで，$\mathcal{N}(x; m, \sigma^2)$ は平均 x，標準偏差 σ のガウス分布を表し，W_V は F0 の変化のしやすさを表すパラメータ (単位:cent) を表す．本研究では，W_V を 100 cent に設定した．

4. 評価実験

本手法の有効性を確認するため，評価実験を行った．

4.1 実験条件

歌声・非歌声 GMM の学習には，「RWC 研究用音楽データベース: ポピュラー音楽 (RWC-MDB-P-2001)」[19] から選んだ 14 歌手 21 曲を用いた．これらの 21 曲に対して，まず，ミックスダウンされたデータとボーカルのみのデータを比較することで，歌声が存在する区間を検出した．次に，歌声 GMM の学習に用いた特徴量は，歌声が存在する区間を，ボーカルのみのデータから推定された F0 を用いて分離した音響信号から計算されたものである．非歌声 GMM は，歌声が存在しない区間を，ミックスダウンされたデータから PreFEst[4] を用いて計算された最も優勢な 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学習した

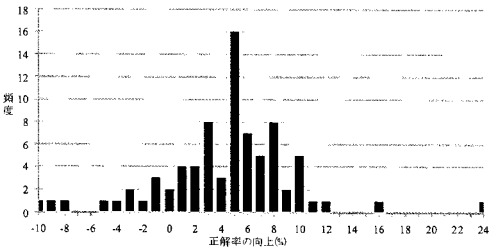


図 5 正解率向上度合いの分布．

評価には，RWC-MDB-P-2001 中の歌声/非歌声 GMM の学習に用いなかった 79 曲を用いた．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[20] 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．

4.2 全体の性能評価

まず，提案手法全体の性能を，音源を考慮しない F0 推定手法である PreFEst と比較し評価した．実験に用いた 79 曲に対する実験結果を図 4 に示す．PreFEst の平均正解率が 75.4%なのに対し，提案手法の平均正解率が 78.8%であることから，本手法を用いることで正解率が 3.4%向上し，誤り率を 13.8%削減できたことがわかる．これにより，特定パートに特化することの効果が確認出来た．図 4 より，多くの楽曲で提案法により正解率が向上する一方，一部の楽曲では逆に正解率が低下している例も見られた．

提案法が有効な楽曲の範囲を調べるため，PreFEst と本手法の正解率の差の分布をグラフにし図 5 に示す．79 曲中 67 曲で提案法により正解率が向上している．最も正解率が向上している#079 の楽曲は，ボーカルパートとピアノパートしか含まない楽曲であり，ピアノが比較的大きな音量でミックスされていた．そのため，PreFEst では伴奏のピアノパートの音高をメロディとして追跡するという誤りが多く発生していた．本手法を用いることで，そのようなピアノパートの音高は歌声確率が低く評価されるため，正しくボーカルパートを追跡出来た．

一方，正解率が低下している#008，#031，#094 等の楽曲では，ボーカルパートにコーラスパートが大きな音量で重なっている例やボーカルが癖の強い声で歌っている例が見られた．これらの楽曲では，分離された歌声と歌声 GMM の学習に用いた歌声が大きく異なっていたため，歌声確率を正しく計算出来なかったのだと考えられる．

その他の楽曲で，推定結果の F0 軌跡を観察すると，比較手法では歌声が徐々に小さくなる箇所で歌声の F0

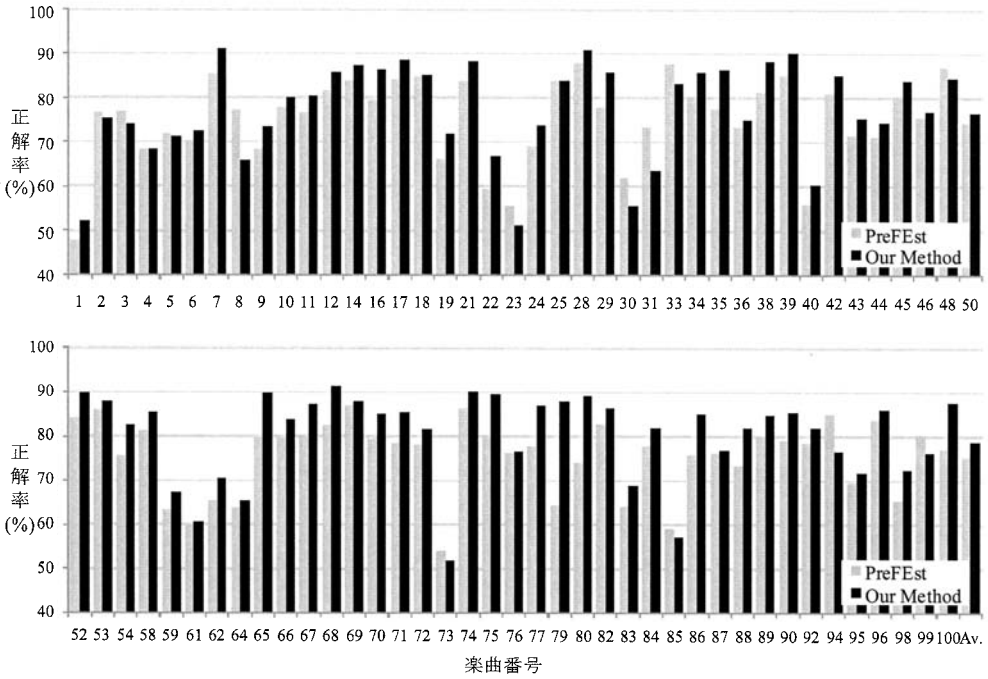


図 4 実験結果: PreFEst と提案法の比較．下段右端の Av. は平均を意味する．

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．比較手法では，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．

図 6 に，実験結果の一部 (#038 の楽曲の 2 分 0 秒から 2 分 10 秒の区間) を図示する．(a) は F0 尤度 (PreFEst における F0 確率密度関数) を図示したものであり，(c) は F0 尤度と歌声確率の積を図示したものである．また，(b) は PreFEst による推定結果を，(d) は本手法による推定結果を表す．(e) は正解として用いたメロディの F0 のアノテーションデータである．図 (a) と (c) の比較することで，歌声確率の導入により，歌声以外の音やノイズの影響で F0 尤度が高くなっている部分が抑制されていることが見て取れる．それに対応して，図 (b) で推定誤りが発生していた区間 (126 秒から 127 秒付近や 121.5 秒付近など) でも，図 (d) では正しく推定されている．

4.3 歌声確率・ビタビ探索の評価

2.4 節で述べたように，本手法の従来の多重 F0 解析法との違いは，歌声確率を重み付けするという点と出力の F0 軌跡を決定する際にビタビ探索を行うという点である．本節では，歌声確率の重み付けとビタビ探索が，それぞれどの程度性能向上に寄与しているかを個別に評価する．歌声確率の効果を評価するため，歌声確率を導入する場合としない場合を比較する．ビタビ探索の評価のために，PreFEst で導入されたマルチエージェントアーキテクチャによる F0 軌跡の追跡手法 (PreFEst-backend と呼ばれる) を比較対象として実験を行う．以下の 4 通りの条件で，実験を行った．

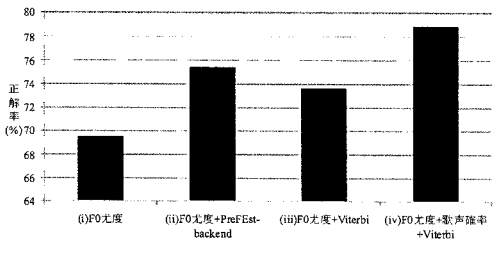


図 7 実験結果: 歌声確率・ビタビ探索の評価．

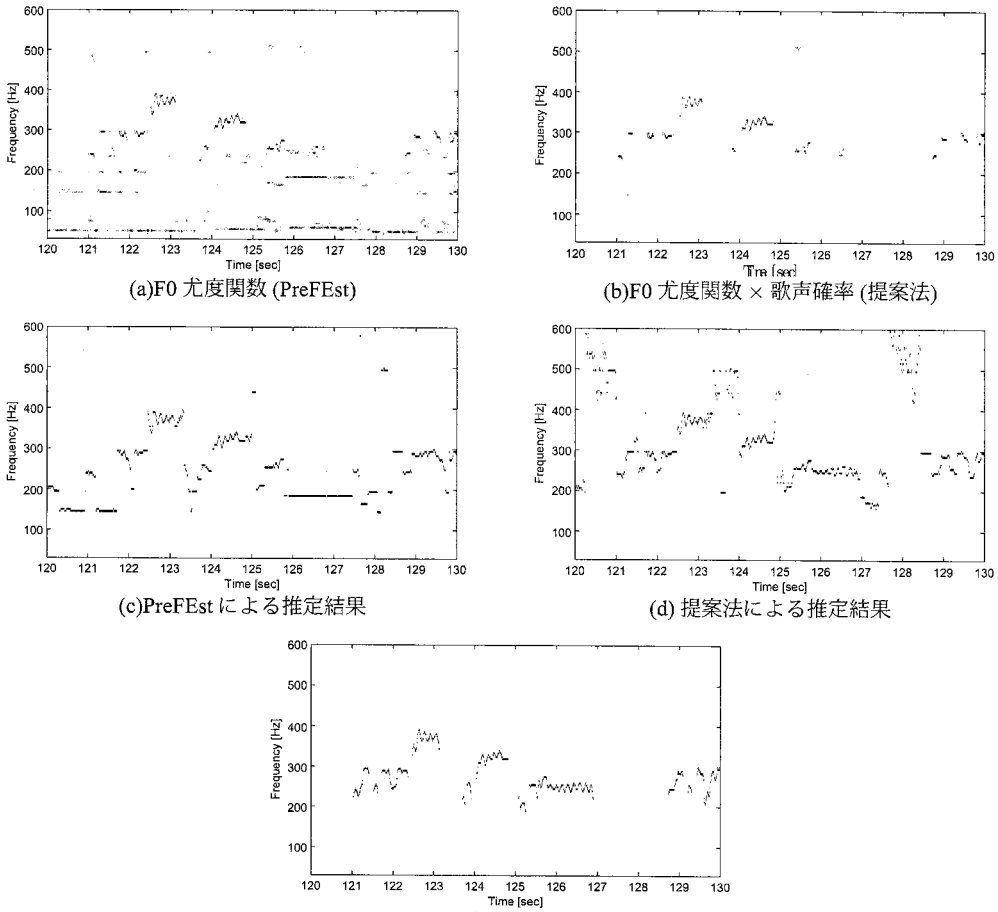
(i) F0 尤度 (PreFEst-core) のみ

(ii) F0 尤度 (PreFEst-core) とマルチエージェントアーキテクチャによる追跡 (PreFEst-backend)

(iii) F0 尤度 (PreFEst-core) とビタビ探索

(iv) F0 尤度 (PreFEst-core) と歌声確率，ビタビ探索 (提案手法)

図 7 に本実験の結果を示す．条件 (iii) と条件 (iv) の比較により，歌声確率導入の純粋な効果が評価出来る．歌声確率の導入により，5.4%精度が向上している．一方，条件 (ii) と条件 (iii) を比較すると，PreFEst-backend は，



(a)F0 尤度関数 (PreFEst)　(b)F0 尤度関数 × 歌声確率 (提案法)

(c)PreFEst による推定結果　(d) 提案法による推定結果

(e) 正解 F0 データ

図 6　実験結果の一例: 楽曲番号 038，2 分 0 秒から 2 分 10 秒まで.

本手法の Viterbi 探索と比較して 1.9%程度性能が良い．シンプルな定式化のビタビ探索で，マルチエージェントアークテクチャを導入し複雑な処理を行う PreFEst-backend に近い性能を得ることが出来たが，今後の性能向上のためには，F0 遷移確率の設計の改良が必要だと考えられる．

4.4 リアルタイム化による性能への影響

本節では，2.3 節で述べたリアルタイム化を行うことで，先読みのフレーム数を変化させることで性能がどのように変化するかについて調べた．これにより，性能がどの程度変化するか，性能を維持するためにはどの程度の先読みが必要かを評価する．

図 8 は，先読みフレーム数と正解率の関係を表す．ただし 1 フレームは 0.01 秒である．図上部の直線は，リアルタイム化しない場合の正解率で，性能の上限と解釈することが出来る．図より，10 フレーム (0.1 秒) 程度先読みを行うことで，リアルタイム化しない場合とほぼ同等の性能が得られることがわかる．また，全く先読みしない場合でも 76.4%と，PreFEst を上回る性能が得られている．このことからも，歌声確率導入の効果が確認できる．

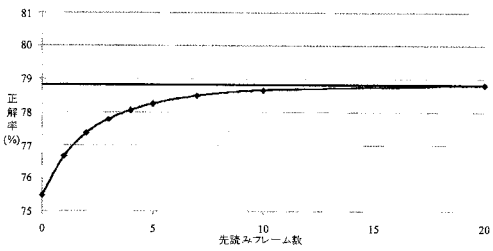


図 8　先読みフレーム数と正解率の関係.

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，特定パートの 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，特定パートに限定することを可能にした．さらに，歌声/非歌声 GMM を用いた歌声確率の計算法と，ビタビ探索による F0 軌跡の探索法によりボーカルパートの F0 推定を実現した．提案手法を用いてボーカルパートに限定することで，F0 の推定精度が 3.4%向上し，本手法の有効性を確認した．

本手法は，複数の音の中から目的の音を聞き分けるという意味で，人間のカクテルパーティ効果の工学的実現であるという点で意義がある．さらに，本研究には，特定パートの F0 推定問題を確率的に定式化し，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．例えば，本研究では F0 尤度の計算に PreFEst を用いたが，この部分をその他の多重 F0 解析手法に置き換えることも可能である．また，F0 遷移確率の設計を改良していくことで，歌声特有の F0 の動きへの対応や音楽的文脈の考慮も行うことが出来る．

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．また，複数の歌手が同時に歌っている場合に対応することも重要な課題である．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．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．

謝辞 本研究の一部は，科研費，21 世紀 COE プログラム，CREST の支援を受けた．また，本研究の実験において，「RWC 研究用音楽データベース：ポピュラー音楽」(RWC-MDB-P-2001)[19] を使用した．最後に，ご討論いただいた亀岡弘和氏 (NTT)，中野倫靖氏 (筑波大学)，北原鉄朗氏 (関西学院大学) に感謝する．

参 考 文 献

1) 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).

2) Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T. and Okuno, H. G.: Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals, *Proceedings of the IEEE International Symposium on Multimedia (ISM2006)*, pp. 257–264 (2006).

3) Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2006)*, Vol. 5, pp. 253–256 (2006).

4) Goto, M.: A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals, *Speech Communication*, Vol. 43, No. 4, pp. 311–329 (2004).

5) Eggink, J. and Brown, G. J.: Gaussian mixture models for extraction of melodic lines from audio recordings, *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR2004)*, pp. 80–83 (2004).

6) Eggink, J. and Brown, G. J.: Extracting melody lines from complex audio, *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR2004)*, pp. 84–91 (2004).

7) Li, Y. and Wang, D.: Detecting Pitch of Singing Voice in Polyphonic Audio, *Proceedigns of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2005)*, Vol. 3, pp. 17–20 (2005).

8) Ryynanen, M. and Klapuri, A.: Transcription of the Singing Melody in Polyphonic Music, *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR2006)*, pp. 222–227 (2006).

9) Poliner, G., Ellis, D., Ehmann, A., Gomez, E., Streich, S. and Ong, B.: Melody Transcription from Music Audio: Approaches and Evaluation, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 4, pp. 1247 – 1256 (2007).

10) Itoyama, K., Kitahara, T., Komatani, K., Ogata, T. and Okuno, H. G.: Automatic Feature Weighting in Automatic Transcription of Specified Part in Polyphonic Music, *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pp. 172–175 (2006).

11) Berenzweig, A. L. and Ellis, D. P. W.: Locating singing voice segments within music signals, *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 119–122 (2001).

12) Tsai, W.-H. and Wang, H.-M.: Automatic Detection and Tracking of Target Singer in Multi-Singer Music Recordings, *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, pp. 221–224 (2004).

13) Nwe, T. L. and Wang, Y.: Automatic Detection of Vocal Segments in Popular Songs, *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pp. 138–145 (2004).

14) Moorer, J. A.: Signal Procesing Aspects of Computer Music: A Survey, *Proceedings of the IEEE*, Vol. 65, No. 8, pp. 1108–1137 (1977).

15) 徳田恵一, 小林隆夫, 今井聖: メル一般化ケプストラムの再帰的計算法, 電子情報通信学会論文誌 A, Vol. J71-A, No. 1, pp. 128–131 (1988).

16) Davis, S. B. and Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition, *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366 (1980).

17) Logan, B.: Mel frequency cepstral coefficients for music modelling, *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, pp. 23–25 (2000).

18) Ohishi, Y., Goto, M., Itou, K. and Takeda, K.: Discrimination between Singing and Speaking Voices, *Proceedings of 9th European Conference on Speech Communication and Technology (Eurospeech 2005)*, pp. 1141–1144 (2005).

19) 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).

20) Goto, M.: AIST Annotation for the RWC Music Database, *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pp. 359–360 (2006).