情報処理学会論文誌　Vol. 49　No. 10　3682–3693 (Oct. 2008)

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

藤原弘将[†1]　後藤真孝[†1]　奥乃博[†2]

本論文では，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有用である．我々は，確率的定式化により，ボーカルパートの F0 推定の問題を音源認識問題（つまり歌声かどうかを認識する問題）と多重ピッチ解析問題に帰着させる．さらに，音源認識問題を歌声・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し，多重ピッチ解析問題を既存手法を拡張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 の系列をビタビ探索によって推定する．評価実験により，歌声区間に対する F0 推定精度が 76.2%から 81.1%に向上し，誤り率を 20.5%削減したことを確認した．

An F0 Estimation Method of Vocal Part in Polyphonic Music by Using Statistical Modelling of Singing Voice and Viterbi Search

Hiromasa Fujihara,[†1] Masataka Goto[†1] and Hiroshi G. Okuno[†2]

This paper describes a method for estimating Fundamental Frequency (F0) of vocal part from polyphonic audio signals. Because melody is performed (sung) by a vocalist in many musical pieces, the estimation of F0s of the vocal part is useful for many applications. We decompose the problem of estimating the vocal F0 into the multiple-F0 estimation problem and the sound source recognition (i.e. estimating a sound source is vocal or not) problem. To deal with the sound source recognition problem, we develop a method of evaluating the vocal probability by using vocal and non-vocal Gaussian mixture models (GMMs). We deal with the multiple-F0 estimation problem by extending the existing method. Finally, we estimate an F0 trajectory that satisfies these stochastic problems by using the Viterbi search. Experimental results show that our method improves estimation accuracy from 76.2% to 81.1%, which is 20.5% reduction of misestimation.

1. はじめに

ポピュラー音楽をはじ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できれば，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同定の研究[1)]と，歌詞を理解するための音楽と歌詞の時間的対応付けの研究[2)]を行ってきた．本論文では，一連の研究の次の段階として，歌詞と並んで楽曲を構成する最も重要な要素の 1 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）推定について述べる．ボーカルパートの F0 推定技術は，歌声を理解するための基礎となる技術であり，文献 1)，2) などでも重要な構成要素となっている．本研究は，これらの歌声理解のための研究の進展に直接貢献することができる．さらに，この技術はボーカルパートの自動採譜やハミング検索，カラオケトラックの自動作成などにも応用することができる．

従来から，多重奏中の単一の F0 を推定する手法は多くあった[3)–8)]．このとき，複数のパートが混在する中で，どのパートの F0 を推定すればよいかという問題が発生する．従来研究では，パートの音源を考慮せず，「メロディ」の F0 推定を目標にしたもの[3)–5),8)]と，推定対象をボーカルパート（歌声パート）に限定したもの[6),7)]の 2 種類がある．音源を考慮しない研究では，具体的には，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ていた．そのため，ボーカルと同時に演奏されるオブリガートなどの F0 を誤って推定してしまうことがあった[3)]．

Li ら[6)]と Ryynänen ら[7)]は多重奏中のボーカルパートの F0 推定の問題に取り組んでいた．文献 6) では，歌声かどうかの判別は行わず，自己相関に基づく方法を用いて高域で最も優勢なピークを選択していた．文献 7) では低レベルの特徴量と，高レベルの音楽的文脈の情報を組み合わせて，ボーカルパートを追跡していた．しかし，特徴量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ボーカルの手がかりとしては不十分であった．このように，従来研究ではボーカルパートを追跡するための手がかりとして音響的特徴

†1 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
†2 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

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．最後に，ビタビ探索を用いた効率的な方法で，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

本章では，与えられた音楽音響信号中のボーカルパートの F0 を推定する手法を説明する．対象とするデータは，市販 CD など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ないと仮定する．

本研究では，ボーカルパート F0 推定の問題を確率的に定式化する．これにより，この問題を，**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できる．F0 尤度の計算は，多重 F0 解析の問題であり，従来手法[⋆1]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，ラプラス分布を用いて設計する．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，多重奏中のボーカルパートの F0 を推定することを可能にした．

2.1 定　式　化

各時刻（フレーム番号）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，

$$F = \{f_t \,|\, t = 1, \cdots, T\} \tag{1}$$

⋆1 本論文の実験では，Goto の PreFEst[3] を用いる．ただし本論文で述べる手法は，PreFEst に特化したものではない．

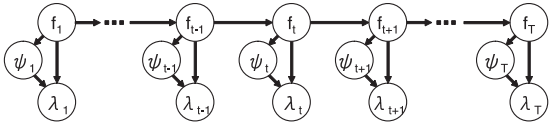


図 1　F，Ψ，Λ の確率的依存関係
Fig. 1　Stochastic dependencies among F, Ψ, Λ.

$$\Psi = \{\psi_t \,|\, t = 1, \cdots, T\} \tag{2}$$

$$\Lambda = \{\lambda_t \,|\, t = 1, \cdots, T\} \tag{3}$$

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，$\lambda_t \in \{s_{\mathrm{V}}, s_{\mathrm{N}}\}$ である．

ボーカルパートの 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，すべ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．

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

ただし，S_V は，すべ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，

$$\hat{F} = \operatorname*{argmax}_F \{\log p(\Lambda = S_V \,|\, F,\ \Psi = O) + \log p(\Psi = O \,|\, F) + \log p(F)\} \tag{5}$$

$$= \operatorname*{argmax}_F \left\{ \sum_{t=1}^{T} \log p(\lambda_t = s_V \,|\, f_t, \psi_t = o_t) + \sum_{t=1}^{T} \log p(\psi_t = o_t \,|\, f_t) + \sum_{t=1}^{T} \log p(f_t \,|\, f_{t-1}) \right\} \tag{6}$$

と分解することができる．

右辺第 1 項 $p(\lambda_t \,|\, f_t, \psi_t)$ は，スペクトル中にある 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，音源認識の問題ととらえることができる．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 解析の問題ととらえることができる．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．このようにして，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，これらの 3 つの条件付き確率を適切に定めることで，F0 推定の際にボーカルパートのみ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 章で述べる．

なお，本論文の定式化は，歌声の基本周波数を知るためには，他の音の基本周波数に関する情報は必要ないという考えかたに基づいている．つまり，観測スペクトルに音源が何個含まれていてもかまわないが，その中で主要な音は 1 つだけと考え，その音の基本周波数を推定することを目指している．逆に考えると，F0 尤度 $p(\psi \mid f)$ とは，推定対象の音が与えられたときに，それ以外の伴奏音などの音をすべて含めた観測スペクトル生成する関数である．これを，本研究では，推定対象の音以外の音の音源数や音源の種類について，すべての可能性を周辺化した状態で，観測スペクトルに F0 がどの程度の強さで含まれているかを表現する関数として解釈する．

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，ビタビ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．

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

これは，音声認識での言語モデルと音響モデル間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

$$\forall f \ \ A(1, f) = \alpha \log p(\lambda_1 \mid f, \psi_1) + \beta \log p(\psi_1 \mid f) \tag{8}$$

(2) 再帰的計算　$(t = 2, \cdots, T)$

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

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

(3) バックトラック

以上で，すべて時刻 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，$B(t, f)$ を後ろ向きにたどっていくことで，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．

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

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

2.3　リアルタイム処理

2.1 節の定式化では，すべ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 を推定することができないので，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，

$$\begin{aligned} \hat{f}_t &= \underset{f_t}{\operatorname{argmax}} \log p(f_t \mid \psi_0 = o_0, \cdots, \psi_{t+N} = o_{t+N}, \lambda_0 = s_{\mathrm{V}}, \cdots, \lambda_{t+N} = s_{\mathrm{V}}) \\ &= \underset{f_t}{\operatorname{argmax}} \log \int \cdots \int p(f_t \cdots f_{t+N} \mid \psi_t = o_t, \\ &\qquad \cdots, \psi_{t+N} = o_{t+N}, \lambda_t = s_{\mathrm{V}}, \cdots, \lambda_{t+N} = s_{\mathrm{V}})\, df_{t+1} \cdots df_{t+N} \end{aligned} \tag{13}$$

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，

$$\hat{f}_t = \underset{f_t}{\operatorname{argmax}} \log p(f_t \cdots f_{t+N} \mid \psi_t = o_t, \cdots, \psi_{t+N} = o_{t+N}, \lambda_t = s_{\mathrm{V}}, \cdots, \lambda_{t+N} = s_{\mathrm{V}}) \tag{14}$$

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できる．

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，本論文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[3)–5),7)] 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

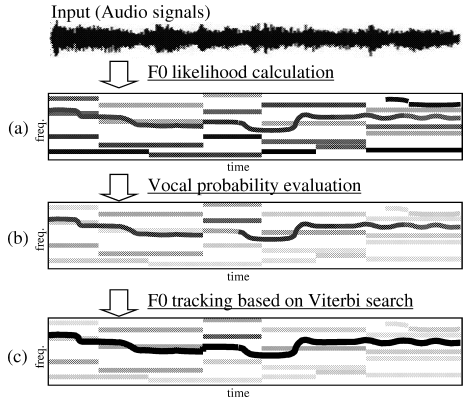


図 2　処理の流れ
Fig. 2　Processing flow.

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できる．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

本章では，歌声・非歌声確率 $p(\lambda_t \,|\, f_t, \psi_t)$，F0 尤度 $p(\psi_t \,|\, f_t)$，F0 遷移確率 $p(f_t \,|\, f_{t-1})$ の具体的な計算方法について述べる．なお，本章の以下の議論では，特に必要な場合を除き，時刻（フレーム番号）を表す添え字 t は省略している．

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 $p(\lambda \,|\, f, \psi)$ 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．つまり，歌声・非歌声確率を計算する問題は，音源が歌声か歌声でないかを推定するという意味で，音源認識の問題ととらえることができる．従来の歌声・非歌声推定手法[9)–11)] はパワーやゼロ交差，MFCC など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できなかった．

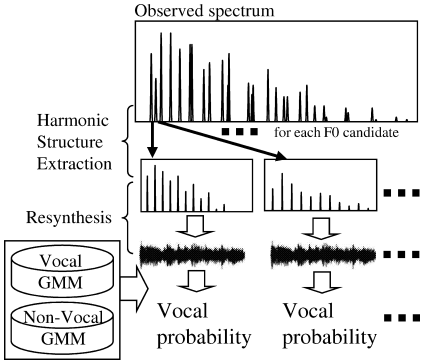


図 3　歌声確率の計算
Fig. 3　Vocal probability calculation.

図 3 に本手法の概要を示す．本研究では，各時刻の観測スペクトル中のすべての F0 候補について，その音が歌声かどうかを計算する．具体的には，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 r 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．h 次倍音（$h = 1, \ldots, H$）の振幅 A_h と周波数 F_h は，以下のように表される．

$$F_h = \operatorname*{argmax}_F \psi(F) \qquad \left(h\overline{F_0} \cdot (1 - 2^{\frac{r}{1200}}) \leq F \leq h\overline{F_0} \cdot (1 + 2^{\frac{r}{1200}})\right) \tag{15}$$

$$A_h = \psi(F_h) \tag{16}$$

ここで，$\psi(F)$ は振幅スペクトルを，$\overline{F_0}$ は抽出したい F0 を表す．本研究では，r を 20 に設定した．

抽出された高調波構造を正弦波重畳モデル[12)] 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 $s(i)$ は，

$$s(i) = \sum_{h=1}^{H} A_h \cos\left(\frac{2\pi F_h}{F_S}\, i\right) \tag{17}$$

と表される．ただし，F_S はサンプリング周波数で，本研究は 16 kHz を用いている．また，i は秒を単位とする時刻を表す．

3.1.2 特徴抽出

歌声確率計算のための GMM で使用する特徴量は，歌声の音響的特徴をよく表現し，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）[13]：
  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）[14] がよく用いられてきた．LPC スペクトルに基づいた LPMCC は，ケプストラムに基づく MFCC と比べてバイアスが小さい[15]．そのため，高い F0 を持つ歌声の特性を表現するのに適していると考えられる．実際，筆者らが以前行った歌手名同定に関する実験では，MFCC と比べて LPMCC を使用した方が歌手名の同定精度が高いという結果が出た[1]．本研究では，LPC スペクトルからメル周波数ケプストラム係数（MFCC）を計算することで LPMCC を抽出した．なお，効率性の観点からも，抽出された高調波構造から信号に戻さずに，スペクトル領域のまま本特徴量を計算する手法の開発が今後の課題となる．
- 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）[16] 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．

$$\Delta f_t = \frac{\sum_{k=-2}^{2} k \cdot f_{t+k}}{\sum_{k=-2}^{2} k^2} \tag{18}$$

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 について，前後の時刻 $t+1$，$t-1$ で接続する F0 の値を求める必要がある．このためには，本来は 2.2 節で述べたビタビ探索と同様の処理が必要である．しかし，この時点では歌声確率 $p(s_{\mathrm{V}} \mid \psi, f)$ は計算できないので，歌声確率を使用せず，3.2 節で述べる F0 尤度 $p(\psi_t \mid f_t)$ と 3.3 節で述べる F0 遷移確率 $p(f_{t+1} \mid f_t)$ のみを用いて以下のように計算する．ある時刻 t の F0 の値 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，

$$f_{t+1} = \operatorname*{argmax}_{f_{t+1}} p(\psi_{t+1} \mid f_{t+1})\, p(f_{t+1} \mid f_t) \tag{19}$$

$$f_{t-1} = \operatorname*{argmax}_{f_{t-1}} p(\psi_{t-1} \mid f_{t-1})\, p(f_t \mid f_{t-1}) \tag{20}$$

である．$t+2$，$t-2$ についても，上記と同様に求める．

3.1.3 歌声・非歌声確率の計算

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．まず，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 と基本周波数 f の同時確率 $p(\psi, f \mid s_{\mathrm{V}})$，$p(\psi, f \mid s_{\mathrm{N}})$ の間の関係を，歌声 GMM の尤度 $\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}; \theta_{\mathrm{V}})$ と非歌声 GMM の尤度 $\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}; \theta_{\mathrm{N}})$ を用いて，

$$\frac{p(\psi, f \mid s_{\mathrm{V}})}{p(\psi, f \mid s_{\mathrm{N}})} = \frac{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{V}})}{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{N}})} \tag{21}$$

と定義する．ただし，$\boldsymbol{x}(\psi, f)$ はあるスペクトル ψ の基本周波数 f の倍音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．ここで，歌声 s_{V} と非歌声 s_{N} の事前分布が等しいことを仮定する．すなわち，

$$p(s_V) = p(s_N) = \frac{1}{2} \tag{22}$$

である．これらを用いて歌声確率 $P(s_{\mathrm{V}} \mid \psi, f)$ は，

$$p(s_{\mathrm{V}} \mid \psi, f) = \frac{p(\psi, f \mid s_{\mathrm{V}}) p(s_{\mathrm{V}})}{\sum_{s=s_{\mathrm{V}},\, s_{\mathrm{N}}} p(\psi, f \mid s)\, p(s)} \tag{23}$$

$$= \frac{p(\psi, f \mid s_{\mathrm{V}})}{p(\psi, f \mid s_{\mathrm{V}}) + p(\psi, f \mid s_{\mathrm{N}})} \tag{24}$$

$$= \frac{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{V}})}{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{V}}) + \mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{N}})} \tag{25}$$

で表される．

なお，$p(\psi, f \,|\, s_{\mathrm{V}})$ と $p(\psi, f \,|\, s_{\mathrm{N}})$ の具体的な関数形は，今回の計算では不要であるが，参考までにその導出について説明する．まず，ψ と f の同時確率 $p(\psi, f)$ は，

$$\begin{aligned} p(\psi, f) &= p(f \,|\, \psi)\, p(\psi) \\ &= p(\psi, f \,|\, s_{\mathrm{V}})\, p(s_{\mathrm{V}}) + p(\psi, f \,|\, s_{\mathrm{N}})\, p(s_{\mathrm{N}}) \end{aligned} \tag{26}$$

と展開できる．一方，式 (21) と式 (22) より，

$$\begin{aligned} &p(\psi, f \,|\, s_{\mathrm{V}})\, p(s_{\mathrm{V}}) + p(\psi, f \,|\, s_{\mathrm{N}})\, p(s_{\mathrm{N}}) \\ &= \frac{1}{2}\, p(\psi, f \,|\, s_{\mathrm{V}}) \left(1 + \frac{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{N}})}{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{V}})} \right) \end{aligned} \tag{27}$$

と展開できる．すると，$p(\psi, f \,|\, s_{\mathrm{V}})$ は，式 (26) と式 (27) より，

$$p(\psi, f \,|\, s_{\mathrm{V}}) = \frac{2 p(f \,|\, \psi)\, p(\psi)}{1 + \frac{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{N}})}{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi, f); \theta_{\mathrm{V}})}} \tag{28}$$

のように表現できる．なお，$p(f \,|\, \psi)$ と $p(\psi)$ は，3.2 節で定義される．$p(\psi, f \,|\, s_{\mathrm{N}})$ についても上記の $p(\psi, f \,|\, s_{\mathrm{V}})$ と同様に導出できる．

3.2 F0 尤 度

F0 尤度の計算には，後藤の PreFEst [3] 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 F0 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 F0 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 $\psi(f)$ が与えられたとき，以後の確率的処理を可能にするため，確率密度関数（PDF）として，

$$p_{\psi}(f) = \frac{\psi(f)}{\int_{-\infty}^{\infty} \psi(f)\, df} \tag{29}$$

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．

$$p(f \,|\, \theta) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w(F)\, p(f \,|\, F)\, dF \tag{30}$$

$$\theta = \{ w(f) \,|\, \mathrm{Fl} \leq f \leq \mathrm{Fh} \} \tag{31}$$

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，

$$\int_{\mathrm{Fh}}^{\mathrm{Fl}} w(f)\, df = 1 \tag{32}$$

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 の確率密度関数を用いて，F0 尤度関数を計算する．まず，確率密度関数 $p(f \,|\, \psi)$ を考え，これを

$$p(f \,|\, \psi) = w(f) \tag{33}$$

と定義する．すると，

$$p(\psi \,|\, f) = \frac{p(f \,|\, \psi)\, p(\psi)}{p(f)} = \frac{p(\psi)}{p(f)}\, w(f) \tag{34}$$

と計算される．ここで，$p(f)$ と $p(\psi)$ がともに一様分布であると仮定し，$\frac{p(\psi)}{p(f)} = C$（C は定数）とおくと，

$$p(\psi \,|\, f) = C w(f) \tag{35}$$

となる．なお，式 (6) の右辺第 2 項に式 (34) を代入した場合，

$$\sum_{t=1}^{T} \log p(\psi_t \,|\, f_t) = T \log C + \sum_{t=1}^{T} \log w(f_t) \tag{36}$$

となり，C は f に無関係な項となるので，C の存在は実際の計算上は無視することができる．

3.3 F0 遷移確率

F0 遷移確率 $p(f_t \,|\, f_{t-1})$ 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

$$p(f_t \,|\, f_{t-1}) = \frac{1}{2b} \exp\!\left(-\frac{|f_t - f_{t-1}|}{b} \right) \tag{37}$$

のように定義する．ただし，f_t，f_{t-1} は cent の単位で表される周波数である[⋆1]．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，$p(f_t \,|\, f_{t-1})$ の標準偏差が 150 cent になるように，

$$b = \sqrt{\frac{150^2}{2}} \tag{38}$$

と規定した．**図 4** に F0 遷移確率を図示する．

F0 遷移確率の分布については，正規分布とも比較したが，予備実験で性能の高かったラ

⋆1 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．

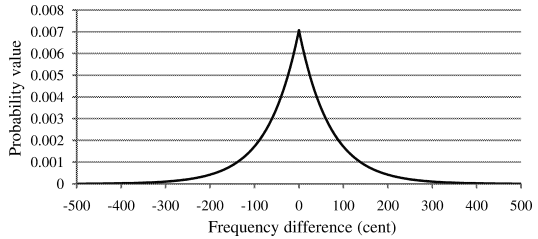


図 4　F0 遷移確率
Fig. 4　F0 transition probability.

プラス分布を採用した．ラプラス分布が適していた理由は，平均付近にピークがありかつ裾が広いという形状が，1 つの音が持続している間は同じような F0 の値が連続し，次の音に遷移する際に大きく変化するという歌声の性質と合致していたためであると考えられる．また，その分散についても，実験的に 150 cent と定めた．なお，F0 遷移確率を学習データから GMM で学習することも試したが，歌声以外の楽器音を誤って追跡する誤りが多く発生し，性能が低下した．このとき，学習によって求めた F0 遷移確率の分布関数は分散が 29.9 cent 程度のラプラス分布に近い急峻な分布となっていた．つまり，歌声以外の楽器音は歌声と比較して F0 の変化が少ないため，学習による分布関数では歌声より歌声以外の楽器音とのあてはまりが良くなってしまったことがこの性能低下の原因と思われる．

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

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）」[17] 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．つまり，100 曲の楽曲を 10 曲ずつの 10 グループに分割し，ある楽曲を評価する際の歌声・非歌声 GMM の学習には，その楽曲が属するグループ以外の 9 グループの楽曲 90 曲を用いた．グループ分けの方法は，100 曲の楽曲番号（No.1 から No.100 まで）を 10 で割った余りが同一のものを同じグループとした．

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[18] を用いた．正解率として，歌声が存在する区間のみを用い，楽曲の歌声が存在する区間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，50 cent 以下とした．なお，4.4 節で述べる実験を除き，2.3 節で述べたリアルタイム処理は行わなかった．

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．

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 と比較し評価する．実験に用いた 100 曲に対する実験結果を**図 5** に示す．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%であることから，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%削減できたことが分かる．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．

最も正解率が向上している No.79 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できた．

その他の楽曲で，推定結果の F0 軌跡を観察すると，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．

図 6 に，実験結果の一部（No.38 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，(b) は F0 尤度と歌声確率の積を図示したものである．また，(c) 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

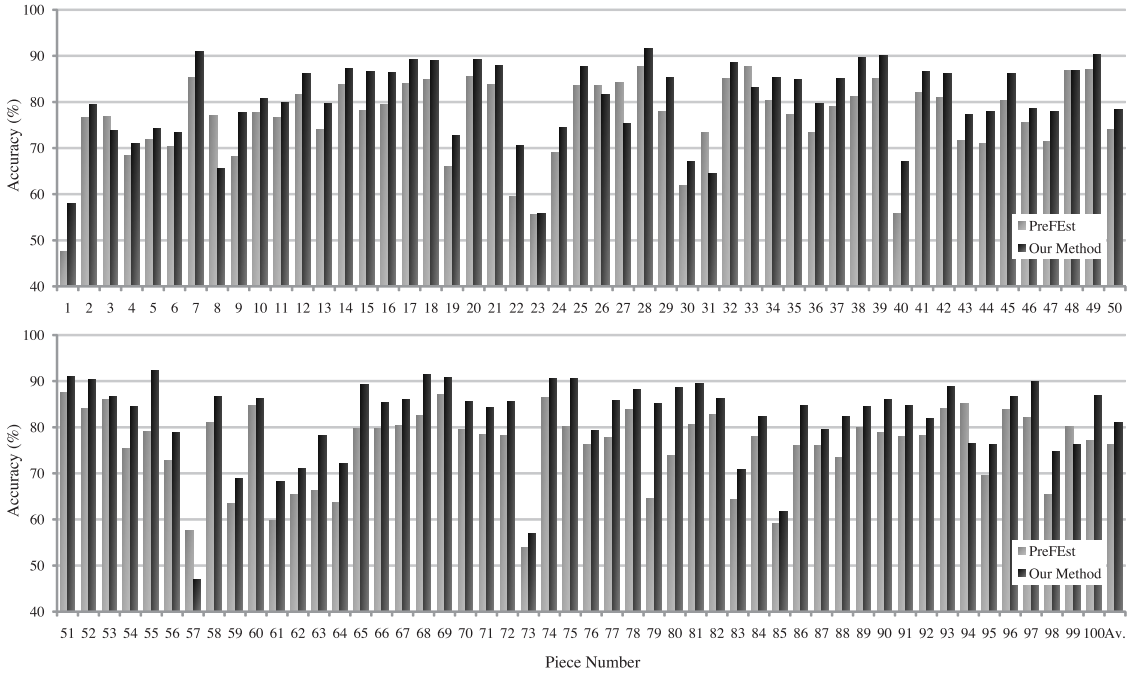


図 5　実験結果：提案法と PreFEst の比較（Av. は平均を意味する）
Fig. 5　Experimental result: The comparison of our method with PreFEst, where Av. means average.

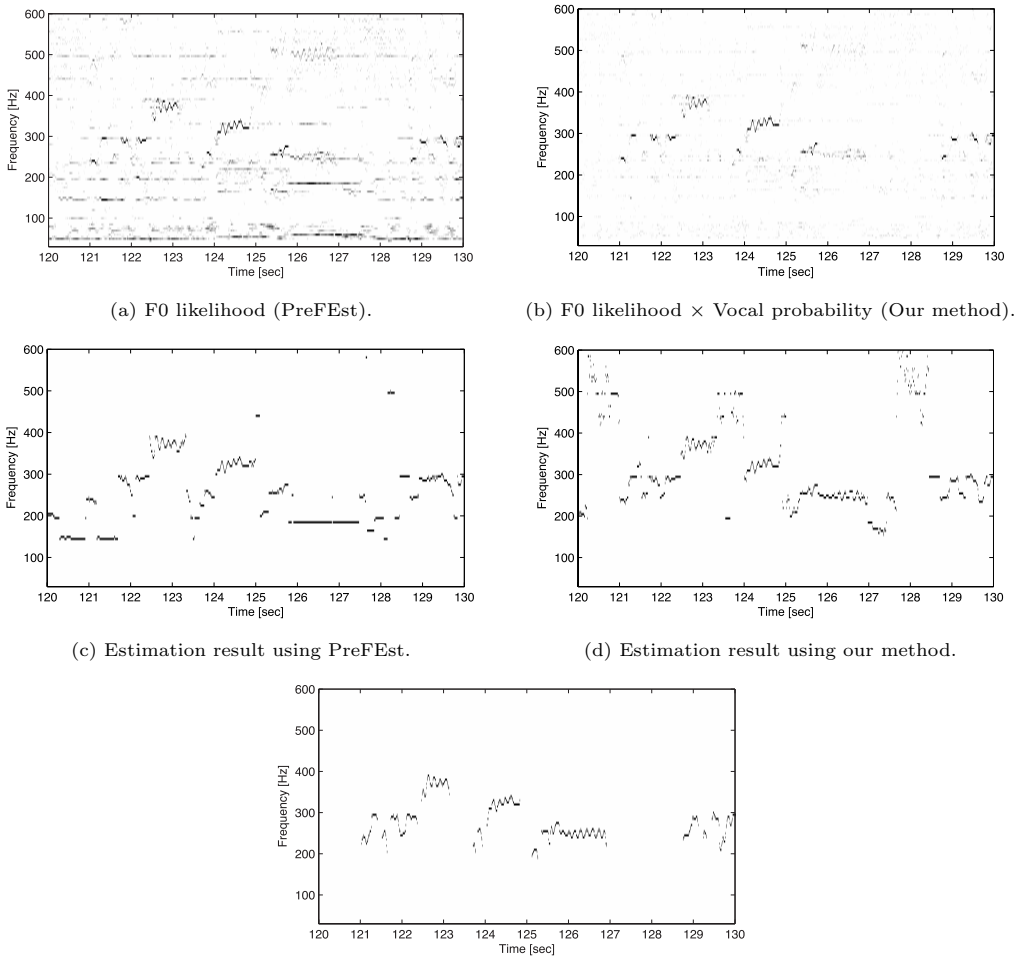
タである．図 (a) と (b) の比較することで，歌声確率の導入により，歌声以外の音やノイズの影響で F0 尤度が高くなっている部分が抑制されていることが見てとれる．それに対応して，図 (c) で推定誤りが発生していた区間（121.5 秒付近や 126 秒から 127 秒付近など）でも，図 (d) では正しく推定されている．

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 7** に示す．100 曲中 90 曲で提案法により正解率が向上している．多くの楽曲で正解率が 3%から 7%程度向上している一方で，逆に正解率が低下している楽曲も一部に存在することが分かる．正解率が低下している楽曲には，ボーカルパートに強いエフェクトがかかっている例やボーカルが癖の強い声で歌っている例，ボーカルの歌唱スタイルがラップである例などがあった．これらの楽曲では，楽曲中の歌声と歌声 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できなかったのだと考えられる．

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 F0 推定手法との違いは，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 で用いら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．

(i)　F0 尤度（PreFEst-core）のみ



(a) F0 likelihood (PreFEst).

(b) F0 likelihood × Vocal probability (Our method).

(c) Estimation result using PreFEst.

(d) Estimation result using our method.

(e) Ground truth.

図 6　実験結果の一例（RWC 研究用音楽データベース RWC-MDB-P-2001 No.38）
Fig. 6　An example of experimental results (RWC Music Database RWC-MDB-P-2001 No.38).

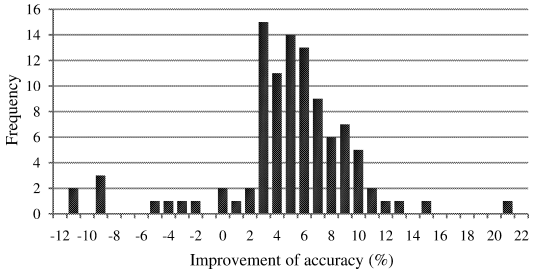


図 7　正解率向上度合いの分布
Fig. 7　Histogram of improvement.

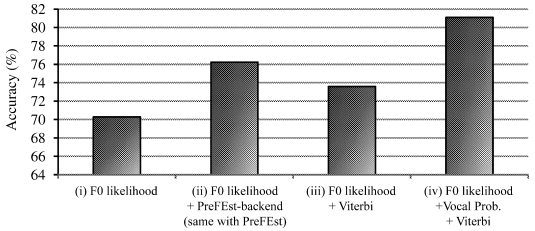


図 8　実験結果：歌声確率・ビタビ探索の評価
Fig. 8　Experimental result: effectiveness of vocal probability and Viterbi search.

(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）
ただし，条件 (ii) は PreFEst [3] と同等である．

図 8 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できる．歌声確率の導入により，7.5 ポイント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，本手法のビタビ探索と比較して約 2.6 ポイント程度性能が良い．本手法のビタビ探索は，マルチエージェントアーキテクチャを導入し複雑な処理を行う PreFEst-backend と比べてシンプルであるという利点があるが，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．

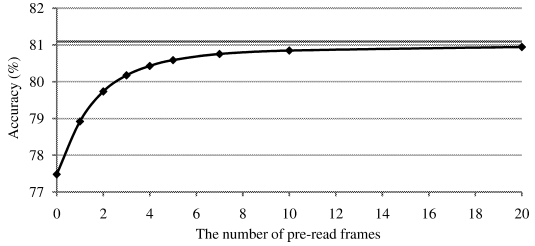


図 9　先読みフレーム数と正解率の関係
Fig. 9　Dependency of accuracy on the number of pre-read frames.

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る．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 9 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できる．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分かる．また，まったく先読みしない場合でも 77.5%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．

5. おわりに

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，従来のメロディ推定の研究と異なり，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，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 4.9 ポイント向上し，本手法の有効性を確認できた．

さらに，本研究には，ボーカルパートの 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．たとえば，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できる．

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．

謝辞　本研究の一部は，科研費，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）[17] 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（産業技術総合研究所），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, *Proc. IEEE International Symposium on Multimedia* (*ISM2006*), pp.257–264 (2006).
3) 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).
4) Marolt, M.: Gaussian Mixture Models for Extraction of Melodic Lines from Audio Recordings, *Proc. 5th International Conference on Music Information Retrieval* (*ISMIR2004*), pp.80–83 (2004).
5) Eggink, J. and Brown, G.J.: Extracting Melody Lines from Complex Audio, *Proc. 5th International Conference on Music Information Retrieval* (*ISMIR2004*), pp.84–91 (2004).
6) Li, Y. and Wang, D.: Detecting Pitch of Singing Voice in Polyphonic Audio, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP2005*), Vol.3, pp.17–20 (2005).
7) Ryynänen, M. and Klapuri, A.: Transcription of the Singing Melody in Polyphonic Music, *Proc. 7th International Conference on Music Information Retrieval* (*ISMIR2006*), pp.222–227 (2006).
8) Poliner, G., Ellis, D., Ehmann, A., Gomez, E., Streich, S. and Ong, B.: Melody Transcription from Music Audio: Approaches and Evaluation, *IEEE Trans. Audio, Speech and Language Processing*, Vol.15, No.4, pp.1247–1256 (2007).
9) Berenzweig, A.L. and Ellis, D.P.W.: Locating Singing Voice Segments within Music Signals, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA2001*), pp.119–122 (2001).
10) Tsai, W.-H. and Wang, H.-M.: Automatic Detection and Tracking of Target Singer in Multi-Singer Music Recordings, *Proc. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2004*), pp.221–224 (2004).
11) Nwe, T.L. and Wang, Y.: Automatic Detection of Vocal Segments in Popular Songs, *Proc. 5th International Conference on Music Information Retrieval* (*ISMIR 2004*), pp.138–145 (2004).
12) Moorer, J.A.: Signal Procesing Aspects of Computer Music: A Survey, *Proc. IEEE*, Vol.65, No.8, pp.1108–1137 (1977).
13) 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).
14) Logan, B.: Mel Frequency Cepstral Coefficients for Music Modelling, *Proc. International Symposium on Music Information Retrieval* (*ISMIR 2000*), pp.23–25 (2000).
15) 今井　聖：音声信号処理音声の性質と聴覚の特性を考慮した信号処理，森北出版株式会社 (1996).
16) Ohishi, Y., Goto, M., Itou, K. and Takeda, K.: Discrimination between Singing and Speaking Voices, *Proc. 9th European Conference on Speech Communication and Technology* (*Eurospeech 2005*), pp.1141–1144 (2005).
17) 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).
18) Goto, M.: AIST Annotation for the RWC Music Database, *Proc. 7th International Conference on Music Information Retrieval* (*ISMIR 2006*), pp.359–360 (2006).

(平成 19 年 12 月 25 日受付)
(平成 20 年 7 月 1 日採録)

藤原　弘将（正会員）

2005 年京都大学工学部情報学科卒業．2007 年同大学大学院情報学研究科知能情報学専攻修士課程修了．同年産業技術総合研究所に入所し，現在に至る．音楽情報処理，音楽情報検索，音声情報処理に興味を持つ．平成 19 年度山下記念研究賞受賞．日本音響学会，電子情報通信学会各会員．

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「情報と知」領域研究員，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．

奥乃　　博（正会員）

1972 年東京大学教養学部基礎科学科卒業．日本電信電話公社，NTT，科学技術振興事業団，東京理科大学を経て，2001 年より京都大学大学院情報学研究科知能情報学専攻教授．博士（工学）．この間，スタンフォード大学客員研究員，東京大学工学部客員助教授．人工知能，音環境理解，ロボット聴覚，音楽情報処理の研究に従事．1990 年度人工知能学会論文賞，IEA/AIE-2001，2005 最優秀論文賞，IEEE/RSJ IROS-2001，2006 Best Paper Nomination Finalist，第 2 回船井情報科学振興賞等受賞．人工知能学会，ACM，IEEE，AAAI 等各会員．