論　文

音楽音響信号を対象としたメロディーとベースの音高推定

後藤　真孝†

F0 Estimation of Melody and Bass Lines in Musical Audio Signals

Masataka GOTO†

あらまし　本論文では，複数の楽器音が混在したモノラルの音楽音響信号に対して，メロディーとベースの音高（基本周波数）を推定する手法を提案する．従来の音高推定手法や音源分離手法は，たかだか三つの音の混合音しか扱うことができず，市販の CD によるジャズやポピュラー音楽の音響信号には有効に機能しなかった．本手法は，混合音下で安定に抽出できない基本周波数成分には依存せず，意図的に制限した周波数帯域（メロディーは中高域，ベースは低域）にある高調波成分が支持する最も優勢な音高を求める．その際，音源数を仮定せずにあらゆる音高の高調波構造が混在しているとみなして混合音をモデル化し，EM（Expectation-Maximization）アルゴリズムにより各高調波構造が相対的にどれくらい優勢かを推定する．更に，マルチエージェントモデルを導入し，各エージェントが音高の時間的な軌跡を追跡することで，最も優勢で安定な音高の軌跡を得ることができる．本手法に基づくシステムを実装して実験した結果，市販の CD からサンプリングした実世界の音響信号に対し，メロディーとベースの音高をリアルタイムに推定できることを確認した．

キーワード　音高推定，ピッチ抽出，音源分離，EM アルゴリズム，音楽理解

1. まえがき

本研究の最終的な目標は，実世界の複雑な音楽音響信号を，人間と同程度まで理解できる処理モデルを実現することである．その第 1 段階として，これまでは音楽音響信号を対象としたビートトラッキングの研究を行ってきた [1]〜[5]．そこでは，まず最初に，音楽的に訓練されていない人でも可能な程度に音楽を理解するモデルを構築し，その後，訓練された音楽家が理解できる程度まで音楽を理解するモデルへと拡張するアプローチの重要性を指摘した [4], [5]．そして，人間は必ずしも音響信号から楽譜に相当するシンボル情報を正確に得て音楽を理解しているわけではないという見地に立って，音響信号を出発点として音符等のシンボルを抽出せずに音楽的な処理を実現してきた．

本研究ではその第 2 段階として，市販の CD（compact disc）などに収録されている，歌声や複数種類の楽器音を同時に含むモノラルの音楽音響信号を対象に，メロディーとベースの音高（本論文では基本周波数の意味で用いる）を推定する処理を実現する．西洋音楽において楽曲の中核を担うメロディーと，調性に密接に関連するベースの理解は，音楽的に訓練された人と訓練されていない人のいずれにとっても基本的な能力であり，それらの音高推定の実現は重要な研究課題である．更に，音高推定の結果は，自動採譜，曲検索のための楽曲情報の自動インデキシング，計算機によるライブ演奏の支援，過去の優れた演奏録音に対する演奏分析，CD を利用したカラオケの伴奏トラックの自動生成等の，様々なアプリケーションにおいて有用である．

多数の音源の音が混ざり合ったモノラルの音響信号中から，ある特定の音源の音高を推定することは，非常に困難な課題である．従来，音高推定手法の多くは，単一音のみか，非周期的な雑音を伴った単一音を収録した音響信号を対象としていた [6]〜[10]．音源分離や自動採譜の研究では，複数の楽器による混合音に対して，各音を分離したり各音に対応するシンボルを得る処理が取り組まれてきたが [11]〜[18]，それらはたかだか三つの楽器音や歌声が同時に鳴る演奏しか扱うことができず，市販の CD による複雑な音響信号に対して音高を推定することはできなかった．CD による音響信号からボーカル音を分離する研究 [19] も報告されているが，事前に与えたボーカル音の楽譜（パート譜）との DP マッチングが不可欠であり，本研究のよ

† 電子技術総合研究所，つくば市
Electrotechnical Laboratory, Tukuba-shi, 305–8568 Japan

うな入力信号の楽譜がない一般的な場合には適用できなかった．このように，複数種類の楽器音や歌声を含む実世界の音楽音響信号に対して，メロディーとベースの音高を推定する手法はまだ実現されていなかった．

本論文では，そのような音響信号に対してメロディーとベースの音高推定を可能にするために，混合音中で最も優勢な音高を推定する手法PreFEst（Predominant-F0 Estimation Method）を提案する．本手法は，各音の高調波構造に対応する確率分布の混合分布（重み付き和）として混合音をモデル化する．その重みの値をEM（Expectation-Maximization）アルゴリズム [20] を用いて推定することで，基本周波数成分の存在に依存せずに，最も優勢な高調波構造を求めることができる．メロディーは中高域において最も優勢な高調波構造をもち，ベースは低域において最も優勢な高調波構造をもつことが多いため，これを意図的に制限した周波数帯域に対して適用すれば，メロディーとベースの音高が推定できる．更に本手法では，マルチエージェントモデルを導入し，音高の時間的な軌跡を追跡しながら安定した推定を実現する．

提案手法をリアルタイムに実行するシステムを分散環境で実装し，CD による音楽音響信号を用いて実験を行った．その結果，単音のメロディーとベースを含む，ポピュラー音楽，ジャズ，クラシックの楽曲 10 曲に対して，メロディーとベースの音高を推定できることを確認した．

2. メロディーとベースの音高推定

本研究では，モノラルの音楽音響信号に対し，その中のメロディーラインとベースラインを推定する問題を解く．メロディーは他よりも際立って聞こえる単音の系列，ベースはアンサンブル中で最も低い単音の系列であり，その時間的な変化の軌跡をそれぞれメロディーライン $D_{\mathrm{m}}(t)$，ベースライン $D_{\mathrm{b}}(t)$ と呼ぶ．時刻 t における基本周波数（F0）を $F_i(t)$ $(i = \mathrm{m}, \mathrm{b})$，振幅を $A_i(t)$ とすると，これらは以下のように表される．

$$D_{\mathrm{m}}(t) = \{F_{\mathrm{m}}(t), A_{\mathrm{m}}(t)\} \quad (1)$$

$$D_{\mathrm{b}}(t) = \{F_{\mathrm{b}}(t), A_{\mathrm{b}}(t)\} \quad (2)$$

つまり，ここでは音符のような楽譜表現にシンボル化することはあえて考えず，基本周波数と振幅の連続値の変化としてメロディーラインとベースラインを求める．

混合音に対して音高推定することが難しい本質的な理由の一つに，時間周波数領域において，ある音の周波数成分が同時に鳴っている他の音の周波数成分と重複することが挙げられる．例えば，歌声，鍵盤楽器（ピアノ等），ギター，ベースギター，ドラムス等で演奏される典型的なポピュラー音楽では，メロディーを担う歌声の高調波構造の一部（特に基本周波数成分）は，鍵盤楽器，ギターの高調波成分やベースギターの高次の高調波成分，スネアドラムの音に含まれるノイズ成分等と頻繁に重複する．そのため，各周波数成分を局所的に追跡するような手法は，複雑な混合音に対しては安定して機能しない．基本周波数成分が存在することを前提に高調波構造を推定する手法もあるが，そのような手法は，ミッシングファンダメンタル（missing fundamental）現象(注 1)を扱えないという大きな欠点をもつ．更に，同時に鳴っている他の音の周波数成分が基本周波数成分と重複すると，有効に機能しない．

メロディーとベースの音高を推定する際の主要な課題は，以上を考慮して次の三つにまとめられる．

（1） 多数の音源の中で，どのようにしてメロディーとベースに着目するか．

（2） 音源数が不明な混合音に対して，どのようにして音高を推定するか．

（3） 音高の候補が複数あるときに，どのようにして適切な音高を選択するか．

本研究では，以下の三つを仮定してこれらを解決する．

- メロディーとベースは高調波構造をもつ．ただし，基本周波数成分の有無は問わない．
- メロディーは中高域において最も優勢な（パワーの大きい）高調波構造をもち，ベースは低域において最も優勢な高調波構造をもつ．
- メロディーとベースの音高は，発音中の時間的な軌跡が連続する傾向をもつ．

以上は多くの場合に当てはまる妥当な仮定である．

各課題に対応する本研究の解決法を以下に示す．

（1） メロディーを求める場合は中高域に，ベースを求める場合は低域に周波数帯域を意図的に制限し，その帯域に含まれる周波数成分が，高調波成分として最も支持するような高調波構造の音高を推定する．その際，その帯域に基本周波数成分が含まれているかど

(注 1)：基本周波数成分が存在しない，あるいは非常に小さい場合でも，高調波成分によって基本周波数に相当する高さが知覚される現象である．メロディー（特に歌声で起きやすい）やベースの音でも，基本周波数成分が非常に小さいことがある．

うかは問わない．

（2） 音源数を仮定せず，対象とするあらゆる音高の高調波構造に対応する確率分布を考え，その混合分布（重み付き和）として観測した周波数成分をモデル化する．そして，その重みの値を EM（Expectation-Maximization）アルゴリズム [20] を用いて推定する．EM アルゴリズムは，隠れ変数を含む確率モデルに対して最ゆう推定を行うための反復アルゴリズムであり，局所最適解を求められる．ここで，最も大きな重みの値をもつ確率分布は，その時点で最も優勢な高調波構造であるとみなせるため，あとはその音高を求めればよい．この手法は基本周波数成分の存在に依存しないため，ミッシングファンダメンタル現象も適切に扱える．

（3） 複数の優勢な音高があるときに，それぞれの時間的な軌跡の連続性を考慮し，最も安定してパワーの大きい軌跡をもつ音高を出力とする．このような軌跡の追跡処理を実現するためにマルチエージェントモデルを導入し，複数のエージェントがそれぞれ異なる音高を追跡することで，安定な音高推定結果を得る．

特に，メロディーの音高を推定する際に，基本周波数付近の帯域では，様々な音の周波数成分が頻繁に重なり合うため，その帯域を意図的に避けることが重要となる．このような基本周波数成分を積極的に用いない手法は，シンギングフォルマント（singing formant）とも関連があると考えられる．シンギングフォルマントとは，男性のオペラ歌唱中の母音が，スペクトル包絡の 2.8 kHz 付近に強いピークをもつ現象である(注2)．オーケストラの大きな伴奏音の平均エネルギー分布は 450 Hz 付近に最大値をもち，その帯域では歌声の基本周波数成分をマスクしてしまう．しかし，シンギングフォルマントがあることによって，高域に存在する歌声の高調波成分が十分優勢となるため，聴衆は歌声を聞き取ることができる [21]．

更に，本研究に関連する音響心理学の知見として，高調波構造をもつ音に関して，文献 [22] では，人間はかなり限定された帯域の情報を主に用いて音の高さの知覚を得ていることが報告され，文献 [23] では，約 1.4 kHz 以下の基本周波数をもつ音の高さの知覚が，基本周波数成分でなく第 2 次以上の高調波成分によって決まることが報告されている．ただし，これらは単一音の知覚に関する報告である．

3. 優勢な音高の推定手法 PreFEst

本研究で提案する，最も優勢な音高を推定する手法の処理の流れを図 1 に示す．まず，入力音響信号に対してマルチレート信号処理を行って瞬時周波数を計算し，瞬時周波数に関連した尺度に基づいて周波数成分の候補を抽出する．次に，2 種類の帯域フィルタ（メロディーライン用とベースライン用）を適用し，それぞれの出力に基づいて，基本周波数（本章では以下，音高ではなく，より正確なこの用語を用いる）の確率密度関数を求める．そして，マルチエージェントモデルを導入し，その確率密度関数の中で有望な各ピークの軌跡を異なるエージェントが追跡して，それぞれの信頼度を評価する．最後に，最も信頼度の高いエージェントがもつ優勢な基本周波数の軌跡を出力する．

3.1 瞬時周波数の算出

本手法では，まず，フィルタバンクの各出力信号に対し，位相の時間微分である瞬時周波数 [24], [25] を計

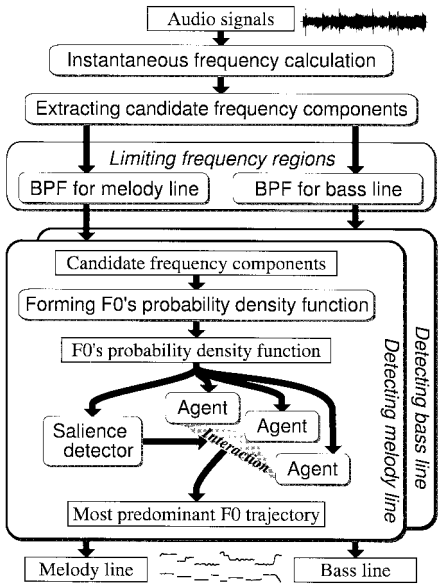

図 1 優勢な音高の推定手法 PreFEst の処理の流れ
Fig. 1 Overview of our predominant-F0 estimation method *PreFEst*.

(注2)：ただし，本研究の適用範囲はオペラ歌唱に限定してはいない．ポピュラー音楽等でも，メロディーの歌声が混合音中で十分聞き取りやすいように通常はミックスダウンがなされており，シンギングフォルマントがなくても，中高域に存在するメロディーの高調波成分が，伴奏に対して十分優勢になっていると考えられる．

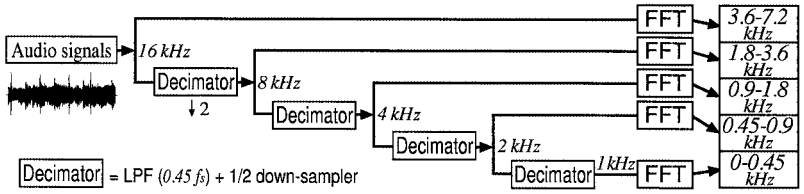

図 2　マルチレートフィルタバンクの構成
Fig. 2　Overview of multirate filter bank.

算する．ここでは，Flanagan の手法 [24] を用い，短時間フーリエ変換（STFT）の出力をフィルタバンク出力と解釈して，効率良く瞬時周波数を計算する．入力音響信号 $x(t)$ に対する窓関数 $h(t)$ を用いた STFT が

$$X(\omega,t) = \int_{-\infty}^{\infty} x(\tau)h(\tau-t)e^{-j\omega\tau}d\tau \qquad (3)$$

$$= a + jb \qquad (4)$$

で定義されるとき，瞬時周波数 $\lambda(\omega,t)$ は次式で求めることができる [24].

$$\lambda(\omega,t) = \omega + \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2+b^2} \qquad (5)$$

ここで窓関数 $h(t)$ として，最適な時間周波数の局所化を与えるガウス関数に 2 階のカーディナル B-スプライン関数を畳み込んで作成した時間窓 [10] を用いる．

この瞬時周波数を計算するのに，単一の STFT のみを用いたのでは，ある周波数帯域における時間分解能や周波数分解能が悪くなってしまう．そこで，マルチレートフィルタバンク [26] を構成し，リアルタイムに実行可能という制約のもとで，ある程度妥当な時間周波数分解能を得る．

設計したバイナリツリー状のフィルタバンクの構成を図 2 に示す．ツリーの各分岐後において，アンチエリアシングフィルタ（FIR 低域フィルタ）と 1/2 ダウンサンプラーによって構成されるデシメータ（decimator）によって，音響信号をダウンサンプリングする．各デシメータの低域フィルタの遮断周波数は 0.45 f_s（f_s は各分岐における標本化周波数）である．現在の実装では，音響信号を標本化周波数 16 kHz，量子化ビット数 16 bit で A–D 変換し，それが最終的に標本化周波数 1 kHz までダウンサンプリングされる．STFT の窓幅は 512 点で，ツリーのそれぞれの葉において時間遅延を補償しながら高速フーリエ変換（FFT）によって計算する．その際，FFT のフレームを 16 kHz において 160 点ずつシフトするため，フレームシフト時間（1 フレームシフト）は 10 ms となる．このフレームシフトを，すべての処理の時間単位とする．

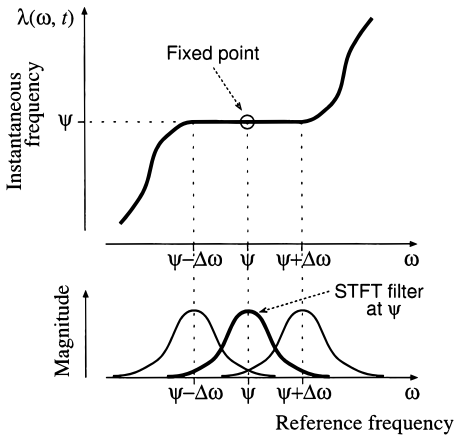

図 3　周波数 ψ の周波数成分がある場合の，STFT フィルタの中心周波数 ω からその出力の瞬時周波数 $\lambda(\omega,t)$ への写像の模式図．上段は周波数成分 ψ に対応する不動点 $(\lambda(\psi,t)-\psi=0)$ の周辺の写像，下段は STFT フィルタの周波数応答を表す
Fig. 3　Sketch of the mapping from the center frequency ω of an STFT filter to the instantaneous frequency $\lambda(\omega,t)$ of its output when there is a frequency component at frequency ψ. The above graph shows the mapping around the fixed point $(\lambda(\psi,t)-\psi=0)$ corresponding to the frequency component ψ and the below graph shows the frequency response of STFT filters.

3.2　周波数成分の候補の抽出

フィルタの中心周波数からその瞬時周波数への写像に基づいて，周波数成分の候補を抽出する [8]〜[10]. まず，ある STFT フィルタの中心周波数 ω からその出力の瞬時周波数 $\lambda(\omega,t)$ への写像を考える．その模式図を図 3 に示す．横軸は，STFT の出力をフィルタバンク出力と解釈した際の各フィルタの中心周波数を

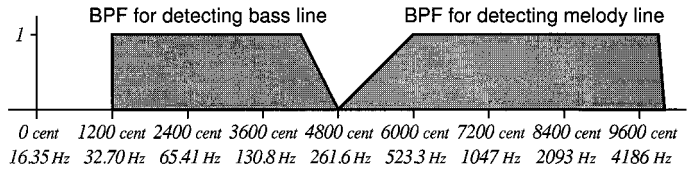

図 4　帯域フィルタ（BPF）の周波数応答
Fig. 4　Frequency responses of bandpass filters (BPFs).

表し，上段の縦軸は，それらのフィルタ出力から求めた瞬時周波数を表す．ここで，図の上段に示したように，もし周波数 ψ の周波数成分があるときには，ψ がこの写像の不動点に位置し，その周辺の瞬時周波数の値はほぼ一定となる [10]．これは，ψ を中心周波数とするフィルタの出力の瞬時周波数が ψ となるだけでなく，その周辺（$\psi \pm \Delta\omega$ の範囲）のフィルタも窓関数で決まる通過帯域内に同じ周波数成分 ψ を含むため，出力の瞬時周波数が ψ となるからである．つまり，全周波数成分の瞬時周波数 $\Psi_f^{(t)}$ は，次式によって抽出することができる [27]．

$$\Psi_f^{(t)} = \{\ \psi \quad | \quad \lambda(\psi,t) - \psi = 0, \ \frac{\partial}{\partial\psi}(\lambda(\psi,t) - \psi) < 0\} \tag{6}$$

これらの周波数成分のパワーは，$\Psi_f^{(t)}$ の各周波数における STFT パワースペクトルの値として得られるため，周波数成分のパワー分布関数 $\Psi_p^{(t)}(\omega)$ を次のように定義できる．

$$\Psi_p^{(t)}(\omega) = \begin{cases} |\,X(\omega,t)\,| & \text{if } \omega \in \Psi_f^{(t)} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

3.3　周波数帯域の制限

抽出した周波数成分に重み付けすることで，周波数帯域を制限する．ここでは，メロディーラインとベースライン用に，2 種類の帯域フィルタ（BPF）を用意する．メロディーライン用の BPF は，典型的なメロディーラインの主要な高調波成分の多くを通過させることができ，かつ，基本周波数付近の重複が頻繁に起きる周波数帯域をある程度遮断できるように設計する．一方，ベースライン用の BPF は，典型的なベースラインの主要な高調波成分の多くを通過させることができ，かつ，他の演奏パートがベースラインよりも優勢になるような周波数帯域をある程度遮断できるように設計する．

現在の実装で用いた BPF の周波数応答を図 4 に示す．本論文では以下，対数スケールの周波数を cent の単位（本来は音高差（音程）を表す尺度）で表し，Hz で表された周波数 f_{Hz} を，次のように cent で表された周波数 f_{cent} に変換する．

$$f_{\mathrm{cent}} = 1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}} \tag{8}$$

平均律の半音は 100 cent に，1 オクターブは 1200 cent に相当する．

周波数 x cent での BPF の周波数応答を $BPF_i(x)$ $(i = \mathrm{m}, \mathrm{b})$ とし，周波数成分のパワー分布関数を $\Psi_p'^{(t)}(x)$ とすると，BPF を通過した周波数成分は $BPF_i(x)\Psi_p'^{(t)}(x)$ と表せる．ただし，$\Psi_p'^{(t)}(x)$ は，周波数軸が cent で表されていることを除けば $\Psi_p^{(t)}(\omega)$ と同じ関数である．ここで，次の段階の準備として，BPF を通過した周波数成分の確率密度関数 $p_\Psi^{(t)}(x)$ を定義する．

$$p_\Psi^{(t)}(x) = \frac{BPF_i(x)\ \Psi_p'^{(t)}(x)}{Pow^{(t)}} \tag{9}$$

$Pow^{(t)}$ は BPF を通過した周波数成分のパワーの合計を表す．

$$Pow^{(t)} = \int_{-\infty}^{\infty} BPF_i(x)\ \Psi_p'^{(t)}(x)\ dx \tag{10}$$

3.4　基本周波数の確率密度関数の推定

それぞれの BPF を通過した周波数成分の候補に対し，各高調波構造が相対的にどれくらい優勢かを表す基本周波数の確率密度関数を求める．そのために本手法では，周波数成分の確率密度関数 $p_\Psi^{(t)}(x)$ が，高調波構造をもつ音をモデル化した確率分布（音モデル）の混合分布モデル（重み付き和のモデル）から生成されたと考える．基本周波数が F の音モデルの確率密度関数を $p(x|F)$ とすると，その混合分布モデル $p(x;\theta^{(t)})$ は次式で定義できる．

$$p(x;\theta^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w^{(t)}(F)\ p(x|F)\ dF \tag{11}$$

$$\theta^{(t)} = \{w^{(t)}(F) \mid \mathrm{Fl}_i \leqq F \leqq \mathrm{Fh}_i\} \tag{12}$$

ここで，Fh_i と Fl_i は，許容される基本周波数の上限と下限であり，$w^{(t)}(F)$ は，次式を満たすような，音モデル $p(x|F)$ の重みである．

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w^{(t)}(F)\, dF = 1 \tag{13}$$

CD 等による実世界の音響信号に対して事前に音源数を仮定することは不可能なため，このように，あらゆる基本周波数の可能性を同時に考慮してモデル化することが重要となる．もし，観測した確率密度関数 $p_\Psi^{(t)}(x)$ がモデル $p(x;\theta^{(t)})$ から生成されたかのようにモデルパラメータ $\theta^{(t)}$ を推定できれば，$p_\Psi^{(t)}(x)$ は個々の音モデルへと分解されたとみなすことができ，その重み $w^{(t)}(F)$ を，基本周波数の確率密度関数 $p_{F0}^{(t)}(F)$ と解釈することができる．

$$p_{F0}^{(t)}(F) = w^{(t)}(F) \quad (\mathrm{Fl}_i \leqq F \leqq \mathrm{Fh}_i) \tag{14}$$

つまり，混合分布中において，ある音モデル $p(x|F)$ が優勢になればなるほど（$w^{(t)}(F)$ が大きくなるほど），$p_{F0}^{(t)}(F)$ において，そのモデルの基本周波数 F の確率が高くなる．

以上から，確率密度関数 $p_\Psi^{(t)}(x)$ を観測したときに，そのモデル $p(x;\theta^{(t)})$ のパラメータ $\theta^{(t)}$ を推定する問題を解けばよいことがわかる．$\theta^{(t)}$ の最ゆう推定量は，次式で定義される平均対数ゆう度を最大化することで得られる．

$$\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ \log p(x;\theta^{(t)})\, dx \tag{15}$$

この最大化問題は解析的に解くことが困難なため，EM（Expectation-Maximization）アルゴリズム [20] を用いて $\theta^{(t)}$ を推定する．EM アルゴリズムは，E ステップ（expectation step）と M ステップ（maximization step）を交互に繰り返し適用することで，不完全な観測データ（この場合，$p_\Psi^{(t)}(x)$）から最ゆう推定を行うための反復アルゴリズムである．ここでは各繰返しにおいて，パラメータ $\theta^{(t)}$ に関して，古いパラメータ推定値 $\theta'^{(t)}$ を更新して新しい（よりもっともらしい）パラメータ推定値 $\overline{\theta^{(t)}}$ を求めていく．$\theta'^{(t)}$ の初期値には，一つ前の時刻 $t-1$ における最終的な推定値を用いる．

周波数 x において観測した各周波数成分が，どの音モデルから生成されたのかを表す隠れ変数（観測できない変数）F を導入して，EM アルゴリズムを以下のように定式化することができる．

（1） E ステップ

平均対数ゆう度の条件付き期待値 $Q(\theta^{(t)}|\theta'^{(t)})$ を計算する．

$$Q\left(\theta^{(t)}|\theta'^{(t)}\right) = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ \mathrm{E}_F\left[\log p\left(x, F;\theta^{(t)}\right)|x;\theta'^{(t)}\right] dx \tag{16}$$

ここで，条件付き期待値 $\mathrm{E}_F[a|b]$ は，条件 b により決定される確率分布をもつ隠れ変数 F に関する，a の期待値を意味する．

（2） M ステップ

$Q(\theta^{(t)}|\theta'^{(t)})$ を $\theta^{(t)}$ の関数として最大化して，更新後の新しい推定値 $\overline{\theta^{(t)}}$ を得る．

$$\overline{\theta^{(t)}} = \underset{\theta^{(t)}}{\mathrm{argmax}}\ Q\left(\theta^{(t)}|\theta'^{(t)}\right) \tag{17}$$

E ステップにおいて，式 (16) より

$$Q\left(\theta^{(t)}|\theta'^{(t)}\right) = \int_{-\infty}^{\infty}\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} p_\Psi^{(t)}(x) p\left(F|x;\theta'^{(t)}\right) \log p\left(x, F;\theta^{(t)}\right) dF dx \tag{18}$$

が得られる．この式中の完全データの対数ゆう度は

$$\log p\left(x, F;\theta^{(t)}\right) = \log\left(w^{(t)}(F)\ p(x|F)\right) \tag{19}$$

で与えられる．次に，M ステップに関しては，式 (17) が，式 (13) を条件とする条件付き変分問題となっている．この問題は，Lagrange の乗数 λ を導入し，次の Euler-Lagrange の微分方程式を用いて解くことができる．

$$\frac{\partial}{\partial w^{(t)}}\left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ p\left(F|x;\theta'^{(t)}\right) \left(\log w^{(t)}(F)\ + \log p(x|F)\right) dx - \lambda\left(w^{(t)}(F) - \tfrac{1}{\mathrm{Fh}_i - \mathrm{Fl}_i}\right)\right) = 0 \tag{20}$$

これより，

$$w^{(t)}(F) = \frac{1}{\lambda}\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ p\left(F|x;\theta'^{(t)}\right)\ dx \tag{21}$$

が得られる．この式において，λ は式 (13) から $\lambda = 1$ と定まり，$p(F|x;\theta'^{(t)})$ はベイズの定理から，

$$p\left(F|x;\theta'^{(t)}\right) = \frac{w'^{(t)}(F)\, p(x|F)}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w'^{(t)}(\eta)\, p(x|\eta)\, d\eta} \qquad (22)$$

となる．ここで，$w'^{(t)}(F)$ は古いパラメータ推定値である（$\theta'^{(t)} = \{w'^{(t)}(F)\}$）．以上から，新しいパラメータ推定値 $\overline{w^{(t)}(F)}$ を求める式は次のようになる．

$$\overline{w^{(t)}(F)} = \int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x) \frac{w'^{(t)}(F)\, p(x|F)}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w'^{(t)}(\eta)\, p(x|\eta)\, d\eta}\, dx \qquad (23)$$

式 (23) を計算するためには，音モデルの確率密度関数 $p(x|F)$ を仮定する必要がある．これは，基本周波数が F のときに，その高調波成分がどの周波数にどれくらい現れるかをモデル化したものである．本研究では，メロディーライン ($i = \mathrm{m}$) とベースライン ($i = \mathrm{b}$) 用に，次のような高調波構造の音モデルを仮定する．

$$p(x|F) = \alpha \sum_{h=1}^{\mathrm{N}_i} c(h)\, G(x; F + 1200 \log_2 h, \mathrm{W}_i) \qquad (24)$$

$$G(x; m, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad (25)$$

ここで，α は正規化係数，N_i は考慮する高調波成分の数（基本周波数成分も数える），W_i^2 はガウス分布 $G(x; m, \sigma)$ の分散を表す．$c(h)$ は，第 h 次高調波成分の振幅を決める関数で，本研究では $c(h) = G(h; 1, \mathrm{H}_i)$（$\mathrm{H}_i$ は定数）とする．この音モデルは，実世界の音響信号中の高調波構造と完全には一致しないが，高調波構造が相対的にどれくらい優勢かを評価する目的においては有効に機能する．ただし，音記憶を導入するなど，今後更に洗練させる余地は残されている．

時刻 t における出力の基本周波数 $F_i(t)$ を決定するには，最も優勢な基本周波数，すなわち基本周波数の確率密度関数 $p_{F0}^{(t)}(F)$（式 (23) を反復計算した最終的な推定値として得られる）を最大にする周波数

$$\underset{F}{\mathrm{argmax}}\ p_{F0}^{(t)}(F) \qquad (26)$$

を求めればよい．しかし，基本周波数の確率密度関数において，同時に鳴っている音の基本周波数に対応する複数のピークが拮抗すると，それらのピークが確率密度関数の最大値として次々に選ばれてしまうことがあるため，このように単純に求めた結果は安定しない．したがって，次節で述べるように，基本周波数に対応するピークの時間的な連続性を考慮する必要がある．

3.5 マルチエージェントモデルによる基本周波数の継時的な追跡

大局的な観点から基本周波数を推定するために，基本周波数の確率密度関数の時間変化において複数のピークの軌跡を継時的に追跡し(注3)，その中で最も優勢で安定した基本周波数の軌跡を選択する．その際，動的に生成・消滅するピークの軌跡を，相互の干渉も考慮しながら，並行して追跡することが不可欠となる．そこで本研究では，そうした動的な追跡処理を柔軟に制御することを可能にする手段として，マルチエージェントモデルを導入する．以前提案したマルチエージェントモデル [28] では，処理中はエージェントの数が固定されていたが，今回のモデルでは，残差駆動型アーキテクチャ [16] のようにエージェントの生成・消滅を動的に行う．

提案するマルチエージェントモデルは，図 5 に示すように，一つの特徴検出器（salience detector）と複数のエージェント（agent）で構成される．エージェントは，追跡周波数の他に，追跡中の軌跡の信頼度と，累積ペナルティを保持し，各時刻において，以下の 5 ステップ（最初の 3 ステップは図 5 と対応）によってこれらを更新していく．

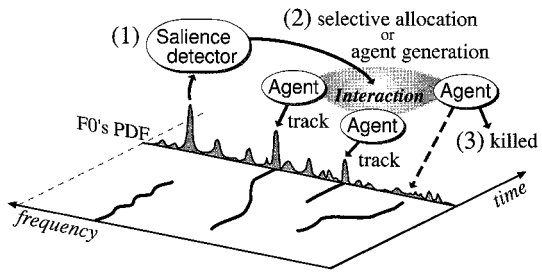

図 5　マルチエージェントモデルによる基本周波数の継時的な追跡
Fig. 5　Sequential F0 tracking by multiple-agent model.

(注 3)：本手法では，音源数を限定せずに基本周波数の確率密度関数を求めているため，適切な音源同定手法を導入することで，複数の音源を同時に追跡する問題へと拡張できる．

（1） 特徴検出器は，基本周波数の確率密度関数（F0's PDF）の中で目立つピーク（最大ピークに応じて動的に変化するしきい値を超えたピーク）を複数検出し，それらがどれくらい有望か（継時的に続いていくか）を評価する(注4)．

（2） エージェント同士が相互作用し，目立つピークをそれに近い軌跡をもつエージェントへと排他的に割り当てる．複数のエージェントが割当て候補に上がる場合には，最も信頼度の高いエージェントへと割り当てる．最も有望で目立つピークが割り当てられなかったときは，そのピークを追跡する新たなエージェントを生成する．

（3） 目立つピークが割り当てられたエージェントの累積ペナルティは，リセットされる．割り当てられなかった場合には一定のペナルティを受け，基本周波数の確率密度関数の中から自分の追跡する次のピークを直接見つけようとする．それも見つからないときには，更にペナルティを受ける．そして，累積ペナルティが一定のしきい値を超えると，そのエージェントは消滅する．

（4） 各エージェントは，割り当てられたピークがどれくらい有望で目立つかに応じて，信頼度を増減する．

（5） 出力の基本周波数 $F_i(t)$ は，信頼度が高く，追跡しているピークの軌跡に沿ったパワーの合計が大きいエージェントに基づいて決定する．振幅 $A_i(t)$ は，基本周波数 $F_i(t)$ の高調波成分を $\Psi_p^{(t)}(\omega)$ から抽出して決定する．

4. システムの実装

音楽音響信号を入力し，推定したメロディーラインとベースラインをリアルタイムに出力するシステムを，提案した手法に基づいて構築した（パラメータの値を表 1 に示す）．出力形式として，視覚化のためのコンピュータグラフィックス，聴覚化のための音響信号，アプリケーションで使用するための連続的に変化する数値（タイムスタンプ付き）の 3 種類に対応した．コンピュータグラフィックスの出力では，時間周波数平面上をスクロールする音高の軌跡を表示するウィンドウと，それと同期してスクロールする周波数成分の候補を表示するウィンドウが提示される（図 6）．音響信号の出力では，検出した $D_i(t)$（式 (1)，式 (2)）に沿って追跡した高調波成分のパワーに基づいて，正弦波重畳モデルを用いて合成する．

表 1 パラメータの値
Table 1 Values of parameters.

$\mathrm{Fh_m} = 9600\,\mathrm{cent}$	$\mathrm{Fh_b} = 4800\,\mathrm{cent}$
$\mathrm{Fl_m} = 3600\,\mathrm{cent}$	$\mathrm{Fl_b} = 1000\,\mathrm{cent}$
$\mathrm{N_m} = 16$	$\mathrm{N_b} = 6$
$\mathrm{W_m} = 17\,\mathrm{cent}$	$\mathrm{W_b} = 17\,\mathrm{cent}$
$\mathrm{H_m} = 5.5$	$\mathrm{H_b} = 2.7$

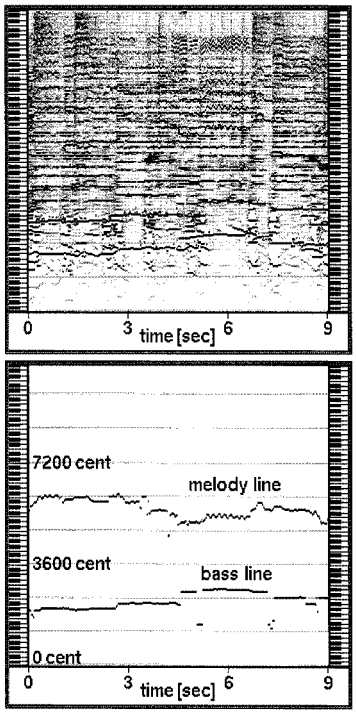

図 6 ウィンドウの画面表示例：ドラムスを伴うポピュラー音楽を入力したときの周波数成分の候補（上）と対応するメロディーラインとベースラインの出力（下）
Fig. 6 Scrolling-window snapshots of candidate frequency components (upper) and the corresponding melody and bass lines detected (lower) for a popular-music excerpt with drum sounds.

本システムを分散環境で実装し，音響信号の入出力，**3.** で提案した手法の計算，中間結果や出力の視覚化といったシステムを構成する各機能を，LAN（Ethernet）上に分散した異なるプロセスとして実行できるようにした．その際，システムの拡張やアプリケーションの開発を容易にするために，RACP（Remote Audio Control Protocol）を設計し，それに基づいて実

(注4)：現在時刻を数フレーム先の時刻とみなし，ピークの軌跡をその時刻まで先読みして追跡することで実現する．

装した．RACP は，RMCP（Remote Music Control Protocol）[29] を音響信号の伝送用に拡張したネットワークプロトコルである．提案手法の計算はパーソナルコンピュータ（Pentium II 450 MHz CPU × 2, Linux 2.2）上で実行され，音響信号の入出力や視覚化の処理はワークステーション（SGI Octane R10000 250 MHz CPU，Irix 6.4）上で実行される．

5. 実験結果

表 2 に示すポピュラー音楽，ジャズ，クラシックの楽曲 10 曲からの抜粋を用いて，システムの動作を確認する実験を行った．入力は市販の CD からサンプリングしたモノラルの音響信号で，それぞれが単音のメロディーと複数種類の楽器音を含んでいる．

システムの出力結果の正誤を判定するために，基準となる正解のメロディーとベースの音高を，人間が手作業で 1 フレーム（10 ms）ごとに指定するための音高情報エディタを開発した．このエディタでは，指定した音高の高調波構造に基づいて正弦波重畳モデルで合成した音や，それを取り除いて残った成分による背景音を聞いたり，それらの周波数成分の推移を時間周波数平面上で見たりしながら作業できる．

こうして作成した正解に基づき，フレームごとに，システムの出力周波数が正解と一致するかどうかを判定して，システムの検出率を求めた．一致して正しいと判定する周波数差の基準は，半音（100 cent）ずれた場合には明らかな誤りであると考え，半音の半分である 50 cent 以下と定めた．ただし，メロディーやベースが鳴っていない区間は評価対象外とした．

5.1 システム全体の検出率の評価実験

システムの検出率を評価した結果を表 2 に示す．各抜粋の多くの部分において，歌声や中域の単音楽器によるメロディーラインと，ベースギターやコントラバスによるベースラインが正しく検出された．メロディーの歌声やソロ楽器が鳴っていない部分では，システムは伴奏音の中に含まれる優勢な音高の軌跡を検出した．これは，提案手法は単に最も優勢な音高を推定するだけで，音源同定までは行っていないためである．

誤検出した箇所では，それまで追跡していたメロディーやベースが鳴り続けているにもかかわらず，一時的にオブリガート等の伴奏パートの方を追跡してしまうことがあった．更に，メロディーの発音直後が十分優勢でないときに，他の伴奏パートの追跡からメロディーの追跡に戻るのが遅れてしまい，その発音直後の軌跡が欠けることがあった．これらの誤検出の本質的な原因は，複数のエージェントが追跡する軌跡の中から適切な軌跡を選ぶ際に，判断の手がかりが不足していることにある．これは，優勢かどうかだけを手がかりとして判断することの限界を示唆しており，今後の研究では，音源同定手法を導入して音源の種類も手がかりに加えることで，対処していく予定である．他の典型的な誤検出は，本来の基本周波数の半分や倍の値を推定してしまう誤りであった．

5.2 マルチエージェントモデルの寄与分の評価実験

表 2 の性能に対し，マルチエージェントモデルがどれくらい寄与しているかを評価するために，マルチエージェントモデルを無効にしたシステム（基本周波数の確率密度関数を最大にする周波数（式 (26)）をそのまま出力するシステム）の検出率を評価した．その結果を **5.1** の結果と比較して表 3 に示す．

表 3 からは，マルチエージェントモデルの導入の効果は小さく，楽曲によっては性能が悪化するような副作用もあるように見える．これは，時間的な連続性を

表 2 メロディーとベースの検出率
Table 2 Detection rates of the melody and bass lines.

タイトル	ジャンル	検出率 [%]	
		メロディー	ベース
Always (Bon Jovi)	ポピュラー	92.4	84.5
Time Goes By (Every Little Thing)	ポピュラー	89.9	64.7
星の降る丘 (Misia)	ポピュラー	89.1	76.6
My Heart Will Go On (Celine Dion)	ポピュラー	88.7	92.2
Spirit of Love (Sing Like Talking)	ポピュラー	85.9	80.0
Vision of Love (Mariah Carey)	ポピュラー	74.5	83.8
Scarborough Fair (Herbie Hancock)	ジャズ	93.6	53.4
On Green Dolphin Street (Miles Davis)	ジャズ	90.8	54.3
Autumn Leaves (Julian “Cannonball” Adderley)	ジャズ	81.2	86.2
Violin Concerto in D, Op. 35 (Tchaikovsky)	クラシック	78.6	77.6

表 3　マルチエージェントモデルの有無による検出率の比較
Table 3　Detection rate comparison: with or without multiple-agent model.

タイトル	マルチエージェントモデルを無効にした場合の検出率 [%]		マルチエージェントモデルを用いた場合の検出率 [%]	
	メロディー	ベース	メロディー	ベース
Always	87.1	79.7	92.4	84.5
Time Goes By	82.6	55.3	89.9	64.7
星の降る丘	86.2	71.3	89.1	76.6
My Heart Will Go On	88.5	89.9	88.7	92.2
Spirit of Love	81.9	64.1	85.9	80.0
Vision of Love	76.1	80.6	74.5	83.8
Scarborough Fair	92.8	52.1	93.6	53.4
On Green Dolphin Street	88.3	54.1	90.8	54.3
Autumn Leaves	87.8	82.5	81.2	86.2
Violin Concerto in D, Op. 35	78.9	71.6	78.6	77.6

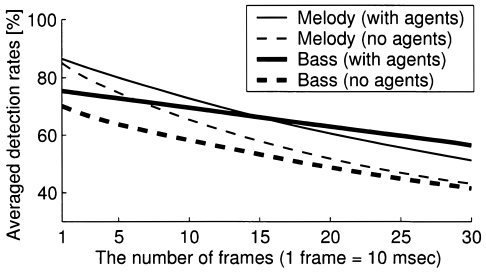

図 7　連続性を考慮して評価した検出率（10 曲の平均値）
Fig. 7　Detection rates evaluated while considering the temporal continuity (averaged rates of 10 songs).

考慮するためにマルチエージェントモデルを導入したにもかかわらず，前述の評価基準は連続性を無視しており，その効果を定量的に評価できていないためである．実際にシステムの出力を観察すると，式 (26) の出力では音高を瞬間的に（1〜2 フレーム）誤ることが頻繁にあるのに対し，マルチエージェントモデルを用いたシステムの出力では，そうした誤りの少ない安定した音高が得られている．

そこで，より的確に比較するために，一定フレーム数の区間連続して正解と一致しなければ正しいと判定しないように，基準を変更して評価した．その比較評価結果を図 7 に示す．横軸が表すフレーム数の区間，常に正解と一致し続けることを条件に検出率を求め，その 10 曲分の平均値を，メロディー（細線）とベース（太線）それぞれについて示した．破線はマルチエージェントモデルを無効にした場合（no agents），実線はマルチエージェントモデルを用いた場合（with agents）の結果である．この図から，マルチエージェントモデルの導入によって性能が確かに向上しており，特にベースの音高推定に効果的であることがわかる．また，**3.5** のマルチエージェントモデルは **3.4** までの後処理として機能しているため，今後の性能向上のためには **3.5** の処理の改良だけでは不十分であり，破線の性能向上（**3.4** までの処理の改良）と破線から実線への改善分の増大（**3.5** の処理の改良）の両者に取り組む必要があるといえる．

6. むすび

本論文では，歌声や複数種類の楽器音を同時に含むモノラルの音響信号に対して，メロディーとベースの音高（基本周波数）を推定する手法 PreFEst について述べた．本手法は，基本周波数成分の有無を問わず，意図的に制限した周波数帯域から得られる部分情報だけを利用して，最も優勢な音高の軌跡を推定できる特長をもつ．音源数を仮定せずに混合音をモデル化し，EM アルゴリズムを適用することで，各高調波構造が相対的にどれくらい優勢かを表す基本周波数の確率密度関数を推定することができた．更に，マルチエージェントモデルを導入することで，基本周波数の時間的な連続性を考慮しながら，最も優勢で安定な音高の軌跡を得ることができた．本手法を実装したシステムを用いて実験した結果，CD による実世界の音響信号中のメロディーとベースの音高を，リアルタイムに推定できることが確認された．

提案手法により推定した基本周波数の確率密度関数は，混合音中の各高調波構造の情報を潜在的に含んでいる．そこで今後は，音源同定手法を導入して，複数の音源の音高を同時に追跡する処理も実現していく予定である．

謝辞　本研究に対し有益な議論をして頂いた，赤穂昭太郎氏，速水悟氏に感謝する．

文　献

[1] 後藤真孝，村岡洋一，“ビートトラッキングシステムの並列計算機への実装 — AP1000 によるリアルタイム音楽情報処理，” 情処学論，vol.37, no.7, pp.1460–1468, 1996.

[2] 後藤真孝，村岡洋一，“音響信号を対象としたリアルタイムビートトラッキングシステム — コード変化検出による打楽器音を含まない音楽への対応，” 信学論（D-II），vol.J81-D-II, no.2, pp.227–237, Feb. 1998.

[3] M. Goto and Y. Muraoka, “Music understanding at the beat level — Real-time beat tracking for audio signals,” in Computational Auditory Scene Analysis, pp.157–176, Lawrence Erlbaum Associates, 1998.

[4] M. Goto and Y. Muraoka, “Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions,” Speech Communication, vol.27, no.3–4, pp.311–335, 1999.

[5] 後藤真孝，音楽音響信号を対象としたリアルタイムビートトラッキングに関する研究，博士論文，早稲田大学理工学部，1998.

[6] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal, “A comparative performance study of several pitch detection algorithms,” IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-24, no.5, pp.399–418, 1976.

[7] A. Nehorai and B. Porat, “Adaptive comb filtering for harmonic signal enhancement,” IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-34, no.5, pp.1124–1138, 1986.

[8] F.J. Charpentier, “Pitch detection using the short-term phase spectrum,” Proc. ICASSP ’86, pp.113–116, 1986.

[9] 阿部敏彦，小林隆夫，今井　聖，“瞬時周波数に基づく雑音環境下でのピッチ推定，” 信学論（D-II），vol.J79-D-II, no.11, pp.1771–1781, Nov. 1996.

[10] 河原英紀，片寄晴弘，R.D. Patterson，A. de Cheveigné，“瞬時周波数を用いた基本周波数の高精度の抽出について，” 音響学音楽音響研資，H-98-116, pp.31–38, 1998.

[11] C. Chafe and D. Jaffe, “Source separation and note identification in polyphonic music,” Proc. ICASSP ’86, pp.1289–1292, 1986.

[12] 片寄晴弘，音楽感性情報処理に関する研究，博士論文，大阪大学基礎工学部，1991.

[13] G.J. Brown and M. Cooke, “Perceptual grouping of musical sounds: A computational model,” J. New Music Research, vol.23, pp.107–132, 1994.

[14] 柏野邦夫，音楽音響信号を対象とする聴覚的情景分析に関する研究，博士論文，東京大学工学部，1994.

[15] 植田　護，橋本周司，“音源分離のためのブラインドデコンポジションアルゴリズム，” 情処学論，vol.38, no.1, pp.146–157, 1997.

[16] 中谷智広，後藤真孝，川端　豪，奥乃　博，“残差駆動型アーキテクチャの提案と音響ストリーム分離への応用，” 人工知能誌，vol.12, no.1, pp.111–120, 1997.

[17] 柏野邦夫，村瀬　洋，“アンサンブル実演奏の自動アンミキサ，” 情処学音楽情報科学研報，98-MUS-24-5, pp.33–40, 1998.

[18] 白土　保，“二重奏音からの基本周波数分離抽出，” 音響誌，vol.54, no.10, pp.715–719, 1998.

[19] 柏野邦夫，村瀬　洋，“パート譜を用いたボーカル音分離システム，” 音講論集，春季 2-9-1, March 1998.

[20] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. Roy. Stat. Soc. B, vol.39, no.1, pp.1–38, 1977.

[21] W. Richards, ed., Natural Computation, The MIT Press, 1988.

[22] R.J. Ritsma, “Frequencies dominant in the perception of the pitch of complex sounds,” J. Acoust. Soc. Am., vol.42, no.1, pp.191–198, 1967.

[23] R. Plomp, “Pitch of complex tones,” J. Acoust. Soc. Am., vol.41, no.6, pp.1526–1533, 1967.

[24] J.L. Flanagan and R.M. Golden, “Phase vocoder,” The Bell System Technical J., vol.45, pp.1493–1509, 1966.

[25] B. Boashash, “Estimating and interpreting the instantaneous frequency of a signal,” Proc. IEEE, vol.80, no.4, pp.520–568, 1992.

[26] M. Vetterli, “A theory of multirate filter banks,” IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-35, no.3, pp.356–372, 1987.

[27] T. Abe, T. Kobayashi, and S. Imai, “The IF spectrogram: A new spectral representation,” Proc. ASVA ’97, pp.423–430, 1997.

[28] M. Goto and Y. Muraoka, “Beat tracking based on multiple-agent architecture — A real-time beat tracking system for audio signals,” Proc. Second Intl. Conf. on Multiagent Systems, pp.103–110, 1996.

[29] 後藤真孝，根山　亮，村岡洋一，“RMCP：遠隔音楽制御用プロトコルを中心とした音楽情報処理，” 情処学論，vol.40, no.3, pp.1335–1345, 1999.

（平成 12 年 1 月 12 日受付，6 月 2 日再受付）

後藤　真孝　（正員）

1993 早大・理工・電子通信卒．1998 同大大学院博士後期課程了．同年，電子技術総合研究所に入所し，現在に至る．博士（工学）．音楽情報処理，音声言語情報処理，マルチモーダルインタラクションなどに興味をもつ．1992 jus 設立 10 周年記念 UNIX 国際シンポジウム論文賞，1993 NICOGRAPH’93 CG 教育シンポジウム最優秀賞，1997 情報処理学会山下記念研究賞，1999 平成 10 年電気関係学会関西支部連合大会奨励賞各受賞．情報処理学会，日本音響学会，日本音楽知覚認知学会，ICMA，ISCA 各会員．