社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－MUS－62 (5)
2005／10／14

ハーモニッククラスタリングと情報量規準による音楽の音高/音源数の推定

亀岡 弘和† 西本 卓也† 嵯峨山茂樹†

† 東京大学大学院情報理工学系研究科
〒113-8656 東京都文京区本郷 7-3-1
E-mail: †{kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

あらまし 本講演では、複数混在する楽音の数、基本周波数、周波数成分の推定するための多重音解析法について述べる。本手法のアイディアの要点は、観測スペクトルをある確率的なメカニズムによって生成された実現値であると考え、その構造を推定することにある。提案法は、汎化誤差を最小にする最小二乗誤差回帰モデルの最良モデル構造を、3 段階処理による MAP 推定と BIC に基づいて探索するものである。そのためにまずこの回帰モデルを、調波構造状の相対配置拘束をもった混合ガウス分布 (これを調波モデルと呼ぶ。) を複数混合したものでモデル化する。これを用いて、ハーモニッククラスタリング、事前分布重みの決定部、BFGS 法により構成される 3 段階処理による MAP 推定で、ある特定のモデル構造における最適モデルパラメータが推定でき、複数のモデル構造に対する「良さ」を BIC に基づき評価し、想定音源数と比例関係にあるパラメータ数を最終的に決定することができる。実演奏音楽信号データに対し、提案法の音高推定性能の評価を行い、高い効果を確認した。
キーワード 多重音解析, ハーモニッククラスタリング, MAP 推定, EM アルゴリズム, BIC

Harmonic Clustering and Information Criterion for Estimating Pitches and the Number of Sources in Music

Hirokazu KAMEOKA†, Takuya NISHIMOTO†, and Shigeki SAGAYAMA†

† Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
E-mail: †{kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

Abstract This paper states a multi-pitch analysis method, including estimation of the number of concurrent sounds, fundamental frequencies and spectral components, based on yet another way of applying statistical learning methods. The idea in this paper is to consider that observed spectrum is a series of realizations from some underlying stochastic generating mechanism and the motivation is to find ways to estimate its structure. The proposed method finds the best structure of least squares regression model with the minimum generalization error based upon a three-stage MAP parameter estimation procedure and a decision making by BIC. The regression model is modeled by multiple harmonically-constrained Gaussian kernel functions, whose maxima are centered over prospective harmonics. The three-stage parameter estimation procedure consists of EM algorithm followed by a determination of the proper prior weights and BFGS gradient search algorithm. The proposed method showed high accuracy in pitch name estimation task of real music performance signal data.
Key words multi-pitch analysis, harmonic clustering, MAP estimation, EM algorithm, BIC

1. はじめに

1 チャネルの多重音解析は、音楽情報処理の分野に留まらず、音声分析や聴覚情景分析などにおいても高い関心がもたれている。音響信号からの MIDI 変換、音楽検索のための音響信号に対する自動インデキシング技術、自動採譜、オーディオ符号化圧縮、音声強調、音声認識のフロントエンドとしての複数話者音声分離などがその効用の代表的な例である。

多重音解析は、何の経験的な情報がない状況では典型的な不良設定問題であり、解析的な解法は一般には存在しない。さらに、音楽や音声のような非定常性は、短時間窓をずらしながら解析するいわゆる短時間分析を余儀なくし、低い周波数分解能や窓関数によるスペクトル漏洩 (拡散) が問題を難しくする。また、予測困難な数々の現象 (混入雑音、残響、ミッシングファンダメンタル、楽音の非調和性) により問題はさらに複雑となる。

これまで、隠れマルコフモデル (HMM) やベイジアンネットワーク (BN) を用いたトップダウンアプローチに基づくさまざま手法が提案された [1]～[4]。このアプローチは、音楽の自動採譜技術の開発を念頭に置き、音名を検出するために観測多重音信号またはスペクトル (スペクトルのピーク周波数) を形成する要素の仮説を立て、仮説の中から最も尤もらしい組合せを探索するものである。この種のアプローチは、音楽のコード進行などの文法的な構造をもつ対象には効果的であり、自動採譜をアプリケーションとする場合に特に有効である。

音楽に限らずさまざまな入力にも汎用に適用できる方法論の開発が望まれるが、パラメトリックモデリングに基づくボトムアップアプローチ [5]～[8] は、自動採譜に特化した技術ではなく、音源分離、音声強調、符号化圧縮などにも広く応用できる可能性をもつ。後藤が提案した PreFEst は、観測パワースペクトルと、あらゆる基本周波数ごとに用意された調波構造状の混合ガウス分布 (調波モデルと呼ぶ。) の重みつき和の分布とが最も近くなるように重み係数を EM アルゴリズムで推定するものであり、推定された各重み係数は、観測スペクトル内に、対応する音高がどれだけ優勢であるかを示す [8]。パワースペクトル領域におけるパラメトリックモデルの最適推定に基づく手法は PreFEst が最初であり、パワースペクトルの加算性により従来までの時間領域での処理に比べてパラメータ最適化の見通しが良いのが特徴である。

ところで、ボトムアップアプローチをとる多くの従来手法では、音高推定精度の向上に主眼が当てられ、音源数は既知である状況やヒューリスティックな閾値により推定するなど、同時発音数の決定法に関しては高い効用にも関わらずそれほど議論の中心として扱われてこなかった [9]～[19] 。その理由は、多重音解析の研究がまだ開始して間もないためであると同時に、

問題自体が本質的に難しいためであると考えられる。例えば、Klapuri は、スペクトル包絡は滑らかであるべきとする先験的な制約に基づき反復的にスペクトルを減算していく手法を提案し、閾値による反復計算の打ち切り判定により音源数を推定するアプローチをとった [19]。このようなアプローチでは、音源数を推定するのにデータに応じて閾値をチューニングする労力が必要になる。こうした状況下で、音源数を自動で推定する方法論は Godsill によって最初に提案された [7]。この手法は、時間領域の重畳信号モデルのパラメータをベイズ推定するものである。時間領域でのモデリングに基づくこの手法では、近似の少ない厳密な目的関数が定義されるが、対象音源数の増加とともにパラメータ最適化が難しくなることが懸念される。

本講演では PreFEst のようにパワースペクトルのパラメトリックモデルを最適推定する枠組に準拠し、音高および音源数を自動推定する手法を提案する。

2. 問題の定式化

2.1 情報量規準のスペクトル解析への応用

本手法のアイディアの要点は、観測スペクトルをある確率的なメカニズムによって生成された実現値であると考え、その構造を推定することにある。我々はこの問題は、情報量規準に基づきあるパラメトリックモデルの最良モデル構造を推定する問題と同型であると考えた。スペクトル解析に情報量規準を適用することを考えた場合、情報量規準の意味と役割は何を確率変数、母集団と考えるかによって大きく違ってくる。

例えば、PreFEst がそうするように、周波数を確率変数、正規化したパワースペクトルを観測された周波数の確率分布 (あるいはヒストグラム) と捉えるのは一つの見方である。この場合、観測パワースペクトルを、調波構造の拘束をもった周波数の確率分布に従う母集団から無作為抽出された周波数標本のヒストグラムであると解釈していることになる。一見して、この考え方に従えば情報量規準によるモデル選択問題にただちに直結するように見受けられるが、ここでは標本数に関する情報が欠落している、あるいは定義できない、という深刻な問題に直面する。なぜなら、周波数標本のデータは確率分布 (実際にはパワースペクトル) の形でしか観測されていないからである。情報量規準の原理によれば、標本数が不明ならばモデルと母集団分布との近さは評価することができず、モデル選択は一般に不可能である。この理由を端的に示す簡単な例を紹介する。例えば、以下のような標本確率分布があったとしよう。

$$f(\omega) = \frac{1}{2}\Big(\delta(\omega-\Omega_0)+\delta(\omega-\Omega_1)\Big),\ \Omega_0 \neq \Omega_1$$

標本確率分布が上式と同じ形であるには、標本データ D が同数の Ω_0 と Ω_1 から構成されている必要があるが、当然その個数には任意性があり、例えば 1 でも ∞ でも良い。今、母集団分布族を混合ガウス分布に限定して議論すると、前者の場合、母集団分布は $(\Omega_0+\Omega_1)/2$ を平均とする単一のガウス分布であると考えるのが合理的であろう。しかし、後者の場合は、Ω_0 と Ω_1 が集中的に生成されていることになるので、母集団分布は Ω_0 と Ω_1 を平均とする 2 つのガウス分布の混合分布が最良のモデル構造と言えそうだ。以上より、標本確率分布が与えられていたとしても、標本数に関する任意性がある限り最良モデル構造を判断することはできない。

従って、情報量規準の原理をスペクトル解析に適用するには、何を確率変数、母集団分布と考えれば良いかをあらためて再考する必要がある。我々が着目したのは、ディジタル信号から得られる観測パワースペクトルが離散系列であるという点である。パワーを確率変数として、離散周波数の各パワースペクトル値に対する回帰分析 (最小二乗誤差推定) を行うという視点に立てば、観測データを有限個の標本と考えることが可能となり、次のような議論ができる。

今、任意の調波構造をモデル化できたと仮定し (これを調波モデルと呼ぶ)、単一音響信号のスペクトルが観測されている状況を考えよう。観測スペクトルに対し、単一の調波モデルを最適推定してもなお生じる誤差は、モデルでは説明し切れない何らかのノイズ (これを以後モデルノイズと呼ぶ。) である。例えば、調波構造をもたないあらゆる音響現象 (背景雑音や雑踏など) に由来するものである。このノイズを、厳密にモデル化するという絶望的な試みは止めて、ある母集団分布からランダムに生成されたものと捉える。さてこの場合、モデルと観測スペクトルとの間の誤差は、ノイズの母集団分布を直接評価するための基準とはなっておらず、モデルの次元に依存したバイアスをもっていると考えるのが情報量基準における重要な思想である。すなわち、モデルと観測との間の誤差をモデルの次元を増やすことで必要以上に小さくすると、かえってモデルの良さは低くなるというのである。情報量規準は、モデルと観測間の誤差の項にそのバイアス分だけの補正項を加えたものであり、観測スペクトルとある個数の調波構造モデルとの誤差が、新たな調波構造モデルで説明するのが合理的なのか、ランダムなモデルノイズと捉えるのが合理的なのかを評価する規準になっている。以上の考え方に基づき、音源数推定問題を情報量規準最小化問題として定式化する。

2.2 モデルと目的関数

このために、まず本節で特定モデル構造 (特定パラメータ数) におけるモデルの推定法について述べる。入力信号の離散パワースペクトル値を正規化した $\{D\} = \{y_i = f(x_i) | i = 1, \cdots, I\}$ を I 個の標本データと考えよう。ただし、x_i は i 番目の離散対数周波数 bin を、y_i は x_i における相対パワー値を表す。このとき、有限データ集合 {D} に対し、最良の回帰モデルを推定するのが当面の目標である。そこで、多重音スペクトルを近似するための関数モデルを $y = h(x;\Theta)$ とする。ただし、説明変数と目的変数 x と y はそれぞれ対数周波数およびパワーに対応する。また、Θ はモデルのパラメータベクトルである。y_i をガウス分布

$$y_i \sim \mathcal{N}\big(h(x_i;\Theta),\sigma^2\big) \equiv p(y_i|\Theta,\sigma^2) \tag{1}$$

に従う、モデル関数 $h(x_i;\Theta)$ とランダムノイズ (モデルノイズ) の和で表される確率変数と仮定すると、ランダムノイズ確率分布モデル $p(\cdot|\Theta,\sigma^2)$ によって与えられる対数尤度

$$\begin{aligned}\ell(\mathbf{y}|\Theta,\sigma^2) &= \log\prod_{i=1}^{I} p(y_i|\Theta,\sigma^2)\\ &= -\frac{I}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\Big(y_i - h(x_i;\Theta)\Big)^2\end{aligned} \tag{2}$$

を最大化すれば良い。これは二乗誤差最小化と同等である。これを拡張し、ベイズの定理より、最大事後確率 (MAP) パラメータ $\hat{\Theta}_{\mathrm{MAP}}$ は、

$$L(\Theta,\sigma^2) \equiv \ell(\mathbf{y}|\Theta,\sigma^2) + \log p(\Theta) \tag{3}$$

を最大化することで得られる。ただし、$p(\Theta)$ は Θ に関する任意の事前確率である。具体的な形については後述する。

次に、回帰モデル $h(x;\Theta)$ の関数形、すなわち多重音パワースペクトルの関数モデルを定式化したい。以後、入力信号に対し、ウェーブレット変換によりスペクトル解析を行う場合を考える。周期信号は周波数領域では線スペクトル状ではなく短時間窓の影響に伴って左右に拡散した分布状となって観測される。単一の周波数成分をガウス分布で近似するならば、$h(x;\Theta)$ は、ガウス分布の重みつき和で表せる。さて、多重音スペクトルは調波構造スペクトルの重ね合わせからなる (パワースペクトルが加算性をもつ) と仮定するならば、PreFEst と同様に調波構造をなすような拘束をもつ N 個のガウス分布の重みつき和 (係数 r_k^n) によって構成される調波モデル (図 1 参照) の K 個分の重みつき和 (係数 w_k)

$$h(x;\Theta) = \sum_{k=1}^{K} w_k \sum_{n=1}^{N} \underbrace{\frac{r_k^n}{\sqrt{2\pi}v_k}\exp\left\{-\frac{(x-\mu_k-\log an)^2}{2{v_k}^2}\right\}}_{h_{kn}(x;\Theta):\text{重み } r_k^n,\ \text{平均 } \mu_k+\log an \text{ のガウス分布}} \tag{4}$$

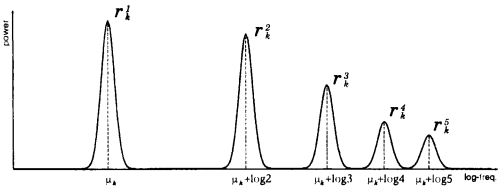

図 1 A single harmonic kernel consists of N number of Gaussian kernels whose maxima are centered over prospective harmonics.

で $h(x;\Theta)$ が表現できる。ただし、μ_k、w_k、ν_k、r_k^n はそれぞれ k 番目の音源の基本周波数、相対パワー、周波数成分の拡散幅、第 n 次倍音の相対パワー (音色) に対応する。a は周波数分析時の周波数分解能に依存して決まる係数である。

2.3 モデル選択

m をある特定モデルにおける自由パラメータ数 (モデル構造の指標) とすると、m は調波モデル数 K に比例した数であり、想定する音源数と (音源数と調波構造の数が対応すると仮定した場合)1 対 1 対応する。赤池、Schwarz によって提唱されたベイズ的情報量規準 (BIC) は

$$\mathrm{BIC}(m) = -2\ell(y|\hat{\Theta}_{\mathrm{MAP}}(m), \sigma^2) + m \log I \qquad (5)$$

で与えられ、モデルの良さを表す規準の 1 つとして知られる [20]〜[22]。ただし、モデル構造 m における $\hat{\Theta}_{\mathrm{MAP}}(m)$ は MAP パラメータを表す。I は標本数をさす。

3. モデルパラメータの MAP 推定

3.1 ハーモニッククラスタリング

式 (3) を最大化するモデルパラメータは解析的には求められないが、勾配法などで数値解析的には求められる。式 (3) の最大化は、ペナルティ関数項 $p(\Theta)$ の下でのモデルと観測スペクトルとの二乗誤差の最小化と解釈できるが、Kullback-Leibler(KL) 尺度の最小化を経由すればパラメータ最適化の見通しが良くなりそうだ。$h(x;\Theta)$ が拘束つきの混合ガウス分布である点を踏まえると、KL 尺度最小化に EM アルゴリズムが適用できそうである。ここでは、調波構造状の相対位置関係をもつ調波モデルの混合分布を観測分布に対して KL 尺度を局所最小化する反復アルゴリズム、[2]ハーモニッククラスタリング を提案する。

w_k と r_k^n が

$$\sum_{\forall k} w_k = 1, \quad {}^{\forall}k, \sum_{\forall n} r_k^n = 1 \qquad (6)$$

を満たすとき、観測スペクトルとモデル間の KL 尺度が定義できるので、本節においてだけ以上を仮定することにする。モデルパラメータに関する両者の KL 尺度の最小化は

$$\ell_{\mathrm{KL}}(y|\Theta) \equiv \sum_{i=1}^{I} f(x_i) \log h(x_i;\Theta) \qquad (7)$$

の最大化と同値である。対数関数は上に凸な関数なので、$\forall x,\ \sum_k \sum_n p(k,n;x) = 1$ を満たす任意の重み関数 $p(k,n;x)$ を用いて Jensen の不等式

$$f(x)\left\langle \log \frac{h_{kn}(x;\Theta)}{f(x)p(k,n;x)} \right\rangle_{p(k,n;x)} \leqq f(x) \log \left\langle \frac{h_{kn}(x;\Theta)}{f(x)p(k,n;x)} \right\rangle_{p(k,n;x)} = \log \frac{h(x;\Theta)}{f(x)} \qquad (8)$$

(注2)：我々の定式化では周波数 x は確率変数ではないので $f(x)$、$h(x;\Theta)$ はいずれも確率分布ではない。そこで、確率則を用いずに EM アルゴリズムと全く同形のアルゴリズムの収束性を示す。実装上は EM アルゴリズムと同値である。

が成立する。ただし、$\langle\cdot\rangle_{p_{k,n}(x)}$ は $p_{k,n}(x)$ による荷重平均演算を表す。これにより、

$$\sum_{\forall k}\sum_{\forall n}\sum_{\forall i} f(x_i)p(k,n;x_i) \log \frac{f(x_i)p(k,n;x_i)}{h_{kn}(x_i;\Theta)} \geqq \sum_{\forall i} f(x_i) \log \frac{f(x_i)}{h(x_i;\Theta)} \qquad (9)$$

なる重要な不等式を得る。この不等式の等号は

$$p(k,n;x_i) = \frac{h_{kn}(x_i;\Theta)}{h(x_i;\Theta)} \qquad (10)$$

のとき成立する。これは、式 (9) の左辺を汎関数 $p(k,n;x)$ に関する変分を 0 と置くことで導かれる。Θ に関する式 (7) の最大化は式 (9) の右辺 (KL 尺度) の最小化と等価であるが、左辺は log の中身が単純にガウス分布なので $p(k,n;x)$ が既知ならば代わりに左辺を最小化するパラメータは解析的に求められそうである。ただし、その場合、変量は $p(k,n;x)$ と Θ の二つになる。$p(k,n;x)$ が式 (10) のとき、本来の目的関数である右辺と、左辺は等号で結ばれる。その状態から左辺を $p(k,n;x)$ を固定のまま Θ に関して最小化すると、両辺ともに必ず減少する。なぜなら式 (9) の不等式より元の目的関数の方が最小化した左辺よりも小さいことが保証されているからである。これを繰り返していけば、下に有界である目的関数を単調減少していくことができ、停留点に収束することができる。以上の考え方は、EM アルゴリズムの収束性に対する別証明になっている。

さて、式 (3) の事前分布項を模擬してここではパラメータに関するペナルティ関数 $\log \tilde{p}(\Theta)$ を導入して目的関数を

$$L_{\mathrm{KL}}(\Theta) \equiv \ell_{\mathrm{KL}}(y|\Theta) + \log \tilde{p}(\Theta) \qquad (11)$$

と設定する。このとき、上述の議論に従えば、補助関数

$$R(\Theta) \equiv \log \tilde{p}(\Theta) + \sum_{\forall k}\sum_{\forall n}\sum_{\forall i} p(k,n;x_i) f(x_i) \log h_{k,n}(x_i;\Theta) \qquad (12)$$

を $p(k,n;x)$ と Θ に関して反復的に最大化すれば良いことが分かる。さて、以後、r_k^n にだけペナルティ関数を導入することを考えよう。r_k^n に何らかのペナルティ関数を仮定するのは、モデルが常識的なスペクトル包絡から極端に逸脱しすぎないような制約を入れることに相当し、モデルの過適応/ 不適応によるオクターブピッチ誤りを防ぐ強い効果がある [8]。ここで、後藤が提案したペナルティ関数

$$\tilde{p}(r_k) \equiv \frac{1}{\beta(d)} \exp\left(-d \sum_{n=1}^{N} \bar{r}_n \log \frac{\bar{r}_n}{r_k^n}\right), \ \sum_{\forall n} \bar{r}_n = 1 \qquad (13)$$

を適用する。r_k^n が $\bar{r}_n$ のときペナルティが最も小さくなることを意味し、d は r_k^n が $\bar{r}_n$ から逸脱したときに与えるペナルティの度合の大きさを、$\beta(d)$ は規格化係数を表す。

以上より、以下の反復計算により、ペナルティ関数の制約下で KL 尺度を最小化することができる。

(1) **初期設定:** $\Theta^{(t)}, (t=0)$ を初期設定する。

(2) **$p(k,n;x)$ の更新 (E-step に相当):** $\Theta^{(t)}$ から式 (10) を計算して (3) に進む。

(3) **Θ の更新 (M-step に相当):** 補助関数 $R(\Theta^{(t+1)})$ を $\Theta^{(t+1)}$ に関して最大化する。t=t+1 として (2) に戻る。

M-step における Θ の各要素パラメータの更新式は

$$\hat{\mu}_k^{(t+1)} = \frac{\sum_{\forall n}\sum_{\forall i} p(k,n|x_i,\Theta^{(t)}) f(x_i)\left(x_i - \log an\right)}{\sum_{\forall n}\sum_{\forall i} p(k,n|x_i,\Theta^{(t)}) f(x_i)}$$

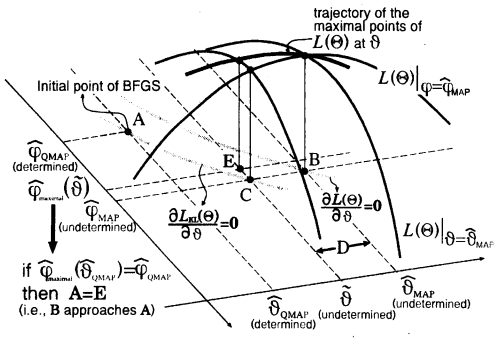

図 2 Hypersurface of $L(\Theta)$. The target solution $\Theta_{MAP} = (\vartheta_{MAP}, \varphi_{MAP})^T$ (point B) is desired to be somewhat close to the starting point: $\Theta_{QMAP} = (\vartheta_{QMAP}, \varphi_{QMAP})^T$ (point A). Considering region of the triangle BCE is significantly smaller than the whole parameter space, we can considerably reduce the search region for the following gradient search method by making point E identical to A.

$$\widehat{w}_k^{(t+1)} = \sum_{\forall n}\sum_{\forall i} p(k,n|x_i,\Theta^{(t)})f(x_i)$$

$$\widehat{r}_k^{n(t+1)} = \frac{d\bar{r}_n + \sum_{\forall i} p(k,n|x_i,\Theta^{(t)})f(x_i)}{d + \sum_{\forall n}\sum_{\forall i} p(k,n|x_i,\Theta^{(t)})f(x_i)}$$

$$\widehat{v}_k^{(t+1)} = \left(\frac{\sum_{\forall n}\sum_{\forall i} p(k,n|x_i,\Theta^{(t)})f(x_i)(x_i-\mu_k^{(t)}-\log a_n)^2}{\sum_{\forall n}\sum_{\forall i} p(k,n|x_i,\Theta^{(t)})f(x_i)} \right)^{\frac{1}{2}}$$

で与えられ、以上の反復アルゴリズムをハーモニッククラスタリングと呼ぶ。ハーモニッククラスタリングと PreFEst の特筆すべき相異点は、調波構造状の拘束つき混合ガウス分布の中心と分散も変数として更新する所にある。PreFEst では、パワースペクトルを、中心と分散が固定の離散配置されたガウス分布の重みだけを推定するため、ある単一周波数のパワー成分分布をその周波数近隣を中心とする複数のガウス分布で近似してしまうことが多く、優勢な調波構造モデルの数と音源数とが必ずしも一致しない。ハーモニッククラスタリングでは分散、中心も自由度をもつ分だけ調波モデルの表現能力 (自由度) が高くなり、一つの観測調波構造が一つの調波モデルで十分近似できるようになる。音源数が調波モデル数を決めることで推定できるため、音源数推定を行う上では明らかにこの性質の方が都合が良い。

目標解 $\widehat{\Theta}_{MAP}$ の表記に倣い、ハーモニッククラスタリング後の収束先パラメータを $\widehat{\Theta}_{QMAP}$ と表記する。$\widehat{\Theta}_{QMAP}$ を初期値として本来の目的関数を最大化するのがこの先の目標である。

3.2 事前分布重みの決定

3.1 節では $\{r_{k,n}\}_{\forall k,\forall n}$ にだけペナルティ関数を導入したので、事前分布を仮定するパラメータとそうでないパラメータをそれぞれ ϑ および φ :

$$\vartheta = \begin{pmatrix} \mu \\ w \\ \nu \end{pmatrix}, \quad \varphi = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_K \end{pmatrix}, \quad \Theta = \begin{pmatrix} \vartheta \\ \varphi \end{pmatrix} \tag{14}$$

$$\mu = (\mu_1, \cdots, \mu_K)^T \quad w = (w_1, \cdots, w_K)^T$$

$$\nu = (\nu_1, \cdots, \nu_K)^T \quad \varphi_k = (r_{k,1}, \cdots, r_{k,N})^T$$

のように区別して表記することにする。

目標解 Θ_{MAP} は、本来、任意に定めることができる事前分布項 $p(\varphi)$ に依存して決まる。しかし、ハーモニッククラスタリングにおいて既にペナルティ関数を特定してしまっているため、$\widehat{\Theta}_{QMAP}$ が本来の目的関数を最大化するための勾配法において良い初期値であるためには事前分布項 $p(\varphi)$ はもはや任意に決めていいものではなくなる。そこで、ここでは、Θ_{QMAP}(図 2 における点 A) が次節で述べる勾配法において良い初期値であるためには $p(\varphi)$ をどのように定めれば良いかを議論する。

以上の目的のため、式 (3) における事前分布項を

$$\log p(\varphi) \equiv \log \prod_{\forall k,n} p(r_{k,n})^{\alpha_{k,n}} \tag{15}$$

$$= \sum_{\forall k,n} \alpha_{kn} \log p(r_{k,n}) \tag{16}$$

のような係数 $\alpha_{k,n}$ の線形結合で定義する。係数 $\alpha_{k,n}$ を事前分布重みと呼ぶ。この定義により、変数 $\alpha_{k,n}$ に応じて任意の ϑ における $L(\Theta)$ の極大点を移動することができるようになる。

事前分布つきパラメータベクトル φ が固定のとき、MAP 推定は最小二乗誤差推定と等価になり、パラメータの MAP 解とハーモニッククラスタリングの収束解はいくらか近いことが予想される。これは、定義は違えど KL 情報量も二乗誤差もいずれもモデルと観測スペクトルとの近さの尺度を表すためである。この仮説が許容されるなら、ある φ において $L(\Theta)$ と $L_{KL}(\Theta)$ の極大点はは常に近いことになる。すなわち、目標解 Θ_{MAP}(図 2 における点 B) と $\varphi = \varphi_{MAP}$ における $L_{KL}(\Theta)$ の極大点 (点 C) が近いことを仮定していることになる。ここで、$\varphi = \widehat{\varphi}_{MAP}$ における $L_{KL}(\Theta)$ の極大点の ϑ のベクトル値を $\widetilde{\vartheta}$ とする。さて、$\widetilde{\vartheta}$ における $L(\Theta)$ の極大点 (点 E) を $(\widetilde{\vartheta}, \widehat{\varphi}_{maximal}(\widetilde{\vartheta}))^T$ と表すと、点 E も点 B にいくらか近いことが期待される。以上より、点 BCE によって囲まれる領域はパラメータ空間全体に比べて相対的に極めて小さいことが予想される。

ここで、変数 $\alpha_{k,n}$ により点 B をいかにして点 A に近づけることができるかを考えたとき、点 E と点 A を一致させることができそうである。このためには、

$$\widetilde{\vartheta} = \widehat{\vartheta}_{QMAP}, \quad \widehat{\varphi}_{maximal}(\widetilde{\vartheta}) = \widehat{\varphi}_{QMAP}$$

$$\Rightarrow \widehat{\varphi}_{maximal}(\widehat{\vartheta}_{QMAP}) = \widehat{\varphi}_{QMAP}$$

$$\Rightarrow \left.\frac{\partial L(\Theta)}{\partial \varphi}\right|_{\Theta=\widehat{\Theta}_{QMAP}} = 0 \tag{17}$$

を満たす必要がある。こうすることで、後述する勾配法では $L(\Theta)$ の極大軌跡に沿って点 E から点 B まで進んでいくだけで良くなる。

さて、簡単のため $p(\varphi)$ を、$\bar{r}_n$ を中心とするガウス分布

$$p(\varphi) \equiv \frac{1}{\sqrt{2\pi\rho_n{}^2}} \exp\left\{ -\frac{(r_{k,n}-\bar{r}_n)^2}{2\rho_n{}^2} \right\} \tag{18}$$

とすると、式 (3) における事前分布項は

$$\log p(\Theta) = \log p(\varphi) = \sum_{\forall k,\forall n} \alpha_k^n \log p(r_{k,n})$$

$$= -\frac{1}{2} \sum_{\forall k}\sum_{\forall n} \alpha_{k,n} \left\{ \log 2\pi\rho_n{}^2 + \frac{(r_{k,n}-\bar{r}_n)^2}{\rho_n{}^2} \right\} \tag{19}$$

で与えられる。ただし、$\rho_n{}^2$ は任意の分散パラメータ (定数) である。式 (17) を満たすためには、$\Theta = \widehat{\Theta}_{QMAP}$ のとき $r_{k,n}$ に関する $L(\Theta, \sigma^2)$ の偏微分

$$\frac{1}{\sigma^2} \sum_{\forall i} \nu_{k,n}(x_i;\vartheta)\Big(f(x_i) - h(x_i;\Theta)\Big) - \alpha_{k,n}\frac{r_{k,n}-\bar{r}_n}{\rho_n{}^2} \tag{20}$$

が 0 である必要がある。ただし、

$$\nu_{k,n}(x_i;\vartheta) = \frac{w_k}{\sqrt{2\pi\nu_k{}^2}} \exp\left\{ -\frac{(x_i-\mu_k-\log a_n)^2}{2\nu_k{}^2} \right\} \tag{21}$$

である。以上より、式 (20) において $\hat{\Theta}_{\mathrm{QMAP}}$ を Θ に代入することにより

$$\hat{\alpha}_{k,n} = \frac{\rho_n{}^2 \sum_{\forall i} \nu_{k,n}(x_i;\hat{\vartheta}_{\mathrm{QMAP}})\left(f(x_i) - h(x_i;\hat{\Theta}_{\mathrm{QMAP}})\right)}{\sigma^2(\hat{r}_{k,n\,\mathrm{QMAP}} - \bar{r}_n)} \tag{22}$$

を得る。

事前分布重みを式 (22) と置くのは、厳密でないまでも目標解をハーモニッククラスタリングの収束解の近くに置くための操作に相当する。ハーモニッククラスタリングと上記の事前分布重み決定に基づく以上の方策は、探索範囲を必ずしも小さくすることを保証するわけではないが、ランダムに設定した初期値から直接勾配法を適用する場合よりも効果的であることが実験的に確認されている。

3.3 BFGS アルゴリズム

以上で得たパラメータ初期点と事前分布重みのもとで、BFGS(Broyden-Fletcher-Goldfarb-Shanno) アルゴリズムによる勾配法を用いて最終推定値を探索する。

t 回目の反復時におけるモデルパラメータベクトルを $\Theta^{(t)}$ とすると、負の目的関数 $Q(\Theta) = -L(\Theta)$ の $\Theta^{(t)}$ の周りの 2 次テーラー展開は

$$Q(\Theta) \approx Q(\Theta^{(t)}) + (\Theta - \Theta^{(t)}) \left.\frac{\partial Q(\Theta)}{\partial \Theta}\right|_{\Theta^{(t)}} + \frac{1}{2}(\Theta - \Theta^{(t)})H(\Theta^{(t)})(\Theta - \Theta^{(t)})^{\mathsf{T}} \tag{23}$$

で与えられる。ただし $H(\Theta^{(t)})$ はヘッセ行列である。BFGS アルゴリズムは、式 (23) で与えられるように $Q(\Theta)$ の最適解の周りを 2 次近似するというニュートン法と同様のアイディアに基づいているが、負定値行列になりうるヘッセ行列の逆行列を直接求めることはせずに、局所的な $Q(\Theta)$ と $\nabla Q(\Theta)$ の情報を使ってこれを正定値制約を満たす行列 $\mathbf{A}$ で逐次近似するのが特徴であり、収束性が保証される。正定値近似行列 $\mathbf{A}$ は、

$$\mathbf{A}_{t+1} = \mathbf{A}_t + \frac{(\nabla Q_{t+1} - \nabla Q_t) \otimes (\nabla Q_{t+1} - \nabla Q_t)}{(\Theta^{(t+1)} - \Theta^{(t)})^{\mathsf{T}}(\nabla Q_{t+1} - \nabla Q_t)} - \frac{\mathbf{A}_t(\Theta^{(t+1)} - \Theta^{(t)}) \otimes \mathbf{A}_t(\Theta^{(t+1)} - \Theta^{(t)})}{(\Theta^{(t+1)} - \Theta^{(t)})^{\mathsf{T}}\mathbf{A}_t(\Theta^{(t+1)} - \Theta^{(t)})} \tag{24}$$

で与えられる有名な BFGS 公式により逐次計算することができる。ただし、$\otimes$ はベクトルの直積演算を表す。

4. モデル選択アルゴリズム

BIC では、モデルの次元が、モデルと誤差の母集団分布の近さ尺度とモデルと観測の近さとのバイアスに相当していることを意味している。このことは、いくつかのモデルがあるデータセットにおいてほぼ等しい対数尤度を持つ場合、低次元のモデルほど母集団分布により近いということを示している。モデルの次元数は調波モデル数に比例するので、母集団分布に最も近いモデル構造を BIC 最小化によって見つけることは音源数推定に対応することになる。BIC 最小モデル構造を探索する処理過程の流れを図 3 に示す。

5. 評価実験

音楽は典型的な多重ピッチ音響信号であることを考慮し、我々は RWC 音楽データベース [23] から選んだ 8 曲の実演奏データに対し、フレームごとの音名推定正解精度を評価した (実験データは表 2 参照)。また、実験条件の詳細を表 1 に示す。音高推定の正解率は、人手でラベル付けした MIDI データ (RWC 音楽データベースの付属データ [23]) を参照用データとして用い、DP マッチングに基づいて自動計算した。この計算においては、置換誤りは脱落誤りと挿入誤りの 2 つの誤りとして見なされるので正解率が負になることもありえる。

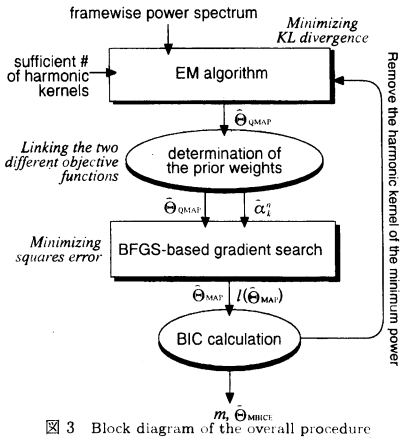

図 3 Block diagram of the overall procedure

表 1 Experimental conditions

spectrum analysis (wavelet transform)	sampling rate/frame shift	16 kHz/32 msec
	frequency resolution	12.0 cent
	frequency range	60–4000 Hz
proposed	initial # of K	10
	σ	3.0×10^{-3}
	ρ_n	$0.01 \times \frac{1}{n}$
PreFEst-core [8]	pitch resolution	20 cent
	# of pitch candidates	200
	variance of Gaussian	9.0
common parameters used in proposed & PreFEst	# of partials	8
	$\bar{r}_n$	$0.6547 \times n^{-2}$
	d	3.0

この評価の目的は、BIC を用いる効果と 3 段階の MAP 推定の効果を明らかにすることにある。まず始めに、BIC に基づいたモデル選択と、推定された強度 $w_k \sum_{\forall n} r_k^n$ に対し閾値処理に基づいて音源数を推定する方法を比較する。次に、多重ピッチ解析法の従来法の 1 つとして広く知られる[3]'PreFEst-core' を比較対象として選んだ。PreFEst は実際は音源数推定について特定の処理を含まないため、適切な比較のため同一条件下で閾値処理により音源数を決定した。以後、評価する 3 種類の手法を以下と呼ぶことにする。**提案法 A**:3 段階 MAP 推定と BIC 最小モデル選択による音源数自動推定処理、**提案法 B**: 3 段階 MAP 推定と閾値よる音源数決定処理、**従来法**: PreFEst-core と閾値による音源数決定処理。これによって BIC と 3 段階 MAP 推定の効果を、提案法 A と提案法 B、および提案法 B と従来法を比較することによってそれぞれ示すことができる。

フレーム毎のピッチ推定結果例を手入力の参照用 MIDI データとともに図 4 に示す。図 5 の結果から、提案法 A は提案法 B よりも優れ、また同様に提案法 B は従来法よりも優れていることが分かる。以上より、我々の提案する 2 つの要素技術である BIC 最小モデル選択による音源数推定法と 3 段階 MAP 推定法は、ともに効果的であることが示せた。

6. まとめと今後の課題

我々は同時発音する音源数とそれぞれの音高とスペクトル成分を求める多重ピッチ分析法を提案した。評価実験では実演奏音楽信号を対象とした音名推定タスクにおいて従来手法より高い性能を示し、また音源数推定における BIC の有用性を示すことができた。今後はこの枠組を複素スペクトル領域に拡張し、重畳信号の 1 チャネルのブラインド音源分離に応用することを検討している。

(注3) : PreFEst は、モデル推定 (core 部) と連続するピッチ軌跡を追跡する処理 (backend 部) で構成されるが、本実験では前者のみを実装した。

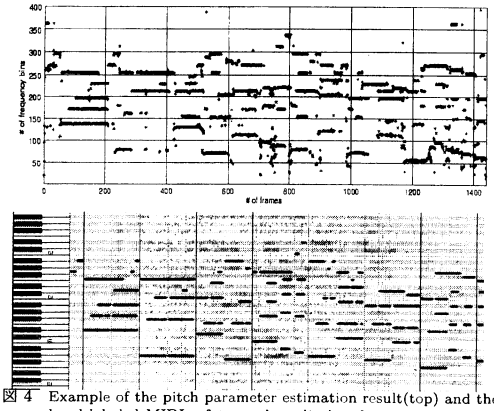

図 4 Example of the pitch parameter estimation result(top) and the hand-labeled MIDI reference data displayed in piano-roll form (bottom).

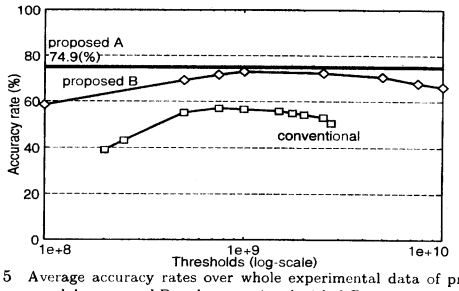

図 5 Average accuracy rates over whole experimental data of proposed A, proposed B and conventional with different thresholds.

表 2 List of the experimental data from RWC music database [23]

Title (Genre)	Instruments	# of frames
Crescent Serenade (Jazz)	Guitar	4427
For Two (Jazz)	Guitar	6555
Jive (Jazz)	Piano	5179
Lounge Away (Jazz)	Guitar	9583
For Two (Jazz)	Piano	9091
Jive (Jazz)	Guitar	3690
Three Gimnopedies no. 1 (Classic)	Piano	6571
Nocturne no.2, op.9-2(Classic)	Piano	7258

謝　辞

本研究に関し、有益な議論をして頂いた後藤真孝 (産総研)、守谷健弘 (NTT CS 研)、武田晴登、齊藤翔一郎 (東大情報理工) の各氏に深く感謝する。

文　献

[1] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka, "Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitive Information Integration Mechanism," *Proc. IJCAI*, Vol. 1, pp. 158–164, 1995.

[2] C. Raphael, "Automatic Transcription of Piano Music," In Proc. International Conference on Music Information Retrieval (ISMIR2002), pp. 15–19, 2002.

[3] A. T. Cemgil, B. Kappen and D. Barber, "Generative Model Based Polyphonic Music Transcription," In Proc. IEEE, Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2003), pp. ?–?, 2003.

[4] R. Leistikow, H. Thornburg, J. Smith III and J. Berger, "Bayesian Identification of Closely-spaced Chords from Single-frame STFT Peaks," In Proc. 7th Int. Conference on Digital Audio Effects (DAFx'04), pp. 228–233, 2004.

[5] M. Feder and E. Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-36, No. 4, pp. 477–489, 1998.

[6] D. Chazan, Y. Stettiner and D. Malah, "Optimal Multi-Pitch Estimation Using the EM Algorithm for Co-Channel Speech Separation," In Proc. IEEE, International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93), Vol. 2, pp. 728–731, 1993.

[7] S. Godsill and M. Davy, "Baysian Harmonic Models for Musical Pitch Estimation and Analysis," In Proc. IEEE, International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002), Vol. 2, pp. 1769–1772, 2002.

[8] M. Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models," Proc. ICASSP2001, Vol. 5, pp. 3365–3368, 2001.

[9] T. W. Parsons, "Separation of Speeh from Interfering Speech by Means of Harmonic Selection," *J. Acoust. Soc. Amer.*, Vol. 60, No. 4, pp. 911–918, 1976.

[10] C. Chafe, D. Jaffe, "Source Separation and Note Identification in Polyphonic Music," In Proc. *ICASSP'86*, pp. 1289–1292, 1986.

[11] H. Katayose, S. Inokuchi, "The Kansei Music System," *Comput. Music J.*, Vol. 13, No. 4, pp. 72-77, 1989.

[12] A. de Cheveigné, "Separation of Concurrent Harmonic Sounds: Fundamental Frequency Estimation and a Time-domain Cancellation Model of Auditory Processing," *J. Acoust. Soc. Amer.*, Vol. 93, No. 6, pp. 3271–3290, 1993.

[13] G. J. Brown, "Computational Auditory Scene Analysis: A Representational Approach," *Ph.D. Thesis, University of Sheffield*, 1992.

[14] T. Nakatani, H. G. Okuno, T. Kawabata, "Residue-driven Architecture for Computational Auditory Scene Analysis," In Proc. *IJCAI-95*, pp. 165–172, 1995.

[15] K. Nishi, S. Ando and S. Aida, "Optimum Harmonics Tracking Filter for Auditory Scene Analysis," In Proc. IEEE, International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96), pp. 573–576. 1996.

[16] A. de Cheveigné, H. Kawahara, "Multiple Period Estimation and Pitch Perception Model," *Speech Comm.*, Vol. 27, No. 3-4, pp. 175–185, 1999.

[17] M. Abe and S. Ando, "Auditory Scene Analysis Based on Time-Frequency Integration of Shared FM and AM (II): Optimum Time-Domain Integration and Stream Sound Reconstruction," *Trans. IEICE*, Vol. J83-D-II, No. 2, pp. 468–477, 2000, (in Japanese).

[18] T. Tolonen, M. Karjalainen, "A Computationally Efficient Multipitch Analysis Model," *IEEE Trans. Speech Audio Process.*, Vol. 8, No. 6, pp. 708–716, 2000.

[19] A. Klapuri, T. Virtanen and J. Holm, "Robust Multipitch Estimation for the Analysis and Manipulation of Polyphonic Musical Signals," In Proc. COST-G6 Conference on Digital Audio Effects (DAFx-00), pp. 141–146, 2000.

[20] H. Akaike, "On Entropy Maximization Principle," In Proc. Applications of Statistics, P. R. Krishnaiah, Ed. Amsterdam, North-Holland, pp. 27–41, 1977.

[21] G. Schwarz, "Estimating the Dimension of a Model," Annals of Statistics. 6, pp. 461–464, 1978.

[22] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," *2nd Inter. Symp. on Information Theory*, Akademia Kiado, Budapest. pp. 267–281, 1973.

[23] 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: クラシック音楽データベースとジャズ音楽データベース," 情報処理学会研究報告, 2002-MUS-44-5, pp. 25–32, 2002.