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

2005－MUS－61 (12)
2005／8／5

調波時間構造化クラスタリング(HTC)による音楽音響特徴量の同時推定

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 本報告では、多重音信号から各音源の音響特徴量 (音高、強度、オンセット、音長、音色など) を同時に推定する新しい方法「調波時間構造化クラスタリング (Harmonic-Temporal structured Clustering: HTC)」を提案する。音楽の音響特徴量抽出は、音楽情報処理分野の中でも近年研究者間で特に関心が高まっている音楽情報検索において極めて重要な要素技術である。HTC は、時間周波数平面に拡散 (漏洩) した観測エネルギーパターン (i.e., パワースペクトルの時系列) を、一つの音源の一連の音響イベントに帰属する個別のエネルギーパターンに分解し、クラスタ化するという考え方に基づいている。このクラスタリングは EM アルゴリズムと同形として理解でき、スペクトルの時間周波数構造モデル (HTM) を複数加算重畳した分布と観測パワースペクトル時系列分布との大域的な近似問題に数学的に等価となる。このモデルをガウス基底関数で構成し、EM アルゴリズムにおけるパラメータ更新式を解析的に導出する。実音楽信号をテストデータとした評価実験で、HTC の高い性能と効果を確認した。
キーワード 音楽音響特徴量抽出, 多重音解析, 調波時間構造化クラスタリング (HTC)

Harmonic-temporal structured clustering (HTC) for simultaneous estimation of audio features in music signal

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 proposes harmonic-temporal structured clustering (HTC) method, that allows simultaneous estimation of pitch, intensity, onset, duration, etc., of each underlying source in multi-stream audio signal, which is expected to be an effective feature extraction for music information retrieval (MIR) systems. HTC decomposes energy diffused in time-frequency space, i.e., a time series of power spectrum, into distinct clusters such that each is originated from a single sound stream. It becomes clear that the problem is equivalent to geometrically approximating the observed time series of power spectrum by superimposed harmonic-temporal structured models (HTMs), whose parameters are directly associated with the acoustic features. The update equations of EM algorithm for the optimal parameter convergence are derived by formulating the model with Gaussian kernel representation. The experiment showed promising results, and verified the potential of the proposed method.
Key words music audio feature extracion, multi-pitch analysis, harmonic-temporal-structured clustering (HTC)

1. はじめに

音楽信号からの音響特徴量抽出は、近年の音楽情報処理分野で最も関心の高いテーマの一つである音楽情報検索の中の重要な要素技術に位置づけられている。本報告では、多重音音楽信号から音響特徴量 (音高、強度、オンセット、音長、音色など) を一挙に推定するための新しい方法論を提案する。

これらの特徴量を適切に抽出する上では、高い性能の多重音解析手法の開発が望まれる。これまで音声分離や自動採譜を目的として提案された数多くの従来研究では、好条件下でしか正しく動作しないものも多く、一般にまだ実用レベルには程遠いと考えられてきた [1]〜[7]。しかしながら、重畳信号モデルのパラメータ推定に基づく手法 [8]、グラフモデルに基づく手法 [9]〜[12]、繰り返しスペクトル減算に基づく手法 [13]、マルチエージェントシステムに基づく手法 [14], [15]、ノンパラメトリックカルマンフィルタに基づく手法 [16], [17]、重畳スペクトルモデルのパラメータ推定に基づく手法 [18]、対数周波数領域における調波構造パターンの逆畳み込み (非線形逆フィルタリング) に基づく手法 [19], [20] に代表される近年の数々の新しいアイデアの登場により、多重音解析は着実に実用化の域に近づ

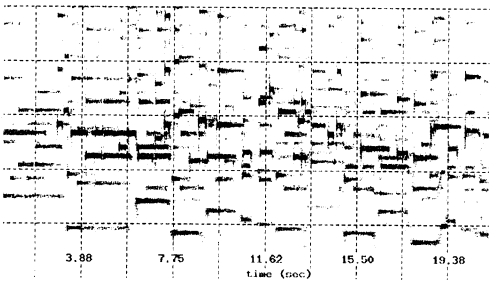


図 1　定義域 D の f(x, t) の濃淡表示の例

いてきた。

これらの手法の多くは、次のような 2 つの独立な手続きをとることで解決を図った。まず周波数次元 (短時間フレーム) における音高構成についての情報を特徴量ベクトル (cf. [9]〜[12])、あるいは尤度ベクトル (cf. [15]〜[18]) という形で抽出し、そのベクトルの時系列から各音高の尤もらしい発音持続区間を隠れマルコフモデル、ベイジアンネットワーク、カルマンフィルタ、マルチエージェントシステムなどで推定する。

複数の旋律で構成される音楽は一般に、同時に複数の音高が、必ずしも同期的でないリズムで発音される多重音響信号であることから、各音の音高とタイミングを解析する問題は、観測パターンが多層な時間周波数構造をもつ点で一般の音響信号処理問題とは大きく異なる性質をもつ。我々は、パワースペクトルは時間に伴って「動的」に変化していくものであるという従来の考え方ではなく、時間周波数平面上に「静的」に配置されたパターンであるという見方から問題を定式化する。この発想により、非同期する多層の動的問題が、2 次元平面上でのスペクトルの「定位」問題に置き換わる。本報告では、時間周波数平面上で生じうるありのままの観測スペクトルパターンをパラメトリックにモデリングし、オーバーオールに最適化する方法論を提案する。提案手法は、時間周波数平面を「情景」と捉える聴覚情景分析 (Computational Auditory Scene Analysis; CASA) の問題に対する一つの新しい解法でもある。

2. 問題の定式化

x を対数周波数、t を時間とし、観測パワースペクトル時系列を f(x, t) とする。ここで、解くべき問題は

$$D = \{x, t \in \mathbb{R} \mid \Omega_0 \leqq x \leqq \Omega_1, T_0 \leqq t \leqq T_1\} \qquad (1)$$

を時間周波数平面上の定義域とする観測密度分布 f(x, t) を、K 個の一連の音響イベントに帰属するエネルギー密度分布 (クラスタ) に分解することである。クラスタリングは「群化」という意味であり、端的には、時間周波数平面上に分布するエネルギーを文字どおり音響イベントごとに群化するのが目標である。

座標 (x, t) における観測エネルギー密度 f(x, t) は、必ずしも単一の音響イベントに完全に帰属するわけではなく、時間および周波数方向のスペクトル漏洩が原因で、複数の音響イベントのエネルギーが複雑に混合 (加算性は一般には成り立たない) されたものとなる。従って、各座標でのエネルギーをただ一つのクラスタが排他的に独占するというより、複数のクラスタで共有するべきである。そこで、音響イベント k にエネルギー密度 f(x, t) を分配するための関数 m(k; x, t) を導入しよう。今、簡単のため、パワーあるいはエネルギーの加算性が成り立つものと仮定すると、エネルギー分配関数 m(k; x, t) は

$$\sum_{\forall k} m(k; x, t) = 1, \quad \forall k, 0 \leqq m(k; x, t) \leqq 1 \qquad (2)$$

を満たせば良く、m(k; x, t)f(x, t) は k 番目の音響イベントのエネルギー密度を表すことになる。これをクラスタ k と呼ぶ。

クラスタリングを定式化するには、クラスタ k の「良し悪し」を評価する何らかの尺度が必要である。そこで、k 番目の音響イベント (発音開始時から減衰するまでをさすこととする。) のエネルギー密度分布のモデルを、パラメータ Θ によって規定される関数 $q_k(x, t; \Theta)$ で表すと、この関数モデルとクラスタ k との間の以下の擬距離

$$\iint_D m(k; x, t) f(x, t) \log \frac{m(k; x, t) f(x, t)}{q_k(x, t; \Theta)} dx dt \qquad (3)$$

は、m(k; x, t)f(x, t) が $q_k(x, t; \Theta)$ と同一な分布のときのみ 0 を与えるため、尺度の一つになる。今、k 番目のクラスタだけではなく、クラスタ全体での良し悪しを評価したいわけだが、以下の条件

$$\iint_D f(x, t) dx dt = \sum_{\forall k} \iint_D q_k(x, t; \Theta) dx dt = 1 \qquad (4)$$

の下では、すべてのクラスタに関する式 (3) の総和

$$J = \sum_{\forall k} \iint_D m(k; x, t) f(x, t) \log \frac{m(k; x, t) f(x, t)}{q_k(x, t; \Theta)} dx dt \qquad (5)$$

はイェンセンの不等式より必ず非負となるので、これが小さいほど、パラメータ Θ とエネルギー分配関数 m(k; x, t), ∀k によって決まるクラスタリングは全体として「良い」ことになる。すなわち、J を式 (2) の下で最小化する m(k; x, t) と Θ を求めることができれば、目標が達成される。さて、式 (5) を、Θ に依存する項としない項に分解し、書き換えると

$$J = -I(\Theta) - \iint_D \lambda(x, t) \left(\sum_{\forall k} m(k; x, t) - 1 \right) dx dt + \sum_{\forall k} \iint_D m(k; x, t) f(x, t) \log m(k; x, t) f(x, t) dx dt$$

$$I(\Theta) \equiv \sum_{\forall k} \iint_D m(k; x, t) f(x, t) \log q_k(x, t; \Theta) dx dt \qquad (6)$$

となる。ただし、第二項は式 (2) を満たすためのラグランジュ未定乗数項である。式 (6) を最小化する Θ と m(k; x, t) は、解析解としては求められないが、k-means アルゴリズムのように Θ と m(k; x, t) を交互に、一方を固定させながら最適化していけば局所最適解に収束させることができる。そこで、まず m(k; x, t) の更新式を変分法により導出する。J の m(k; x, t) に関する 1 次変分は

$$\frac{\delta J}{\delta m_k} = \iint_D \frac{\partial J}{\partial m_k} \delta m_k dx dt \qquad (7)$$

で与えられるので、

$$\frac{\partial J}{\partial m_k} = f(x,t)\left(1 + \log\frac{m(k;x,t)}{q_k(x,t;\Theta)}\right) - \lambda(x,t) \qquad (8)$$

を 0 と置くと、

$$m(k;x,t) = q_k(x,t;\Theta)\exp\left(\frac{\lambda(x,t)}{f(x,t)} - 1\right) \qquad (9)$$

を得る。式 (2)、(9) よりラグランジュ未定乗数 $\lambda(x,t)$ は

$$\lambda(x,t) = f(x,t)\left(1 - \log\sum_{\forall k} q_k(x,t;\Theta)\right) \qquad (10)$$

となるので、式 (10) を式 (9) に代入すると

$$\hat{m}(k;x,t) = \frac{q_k(x,t;\Theta)}{\sum_{\forall k} q_k(x,t;\Theta)} \qquad (11)$$

が得られる。以上で、Θ が固定であるときの最適エネルギー分配関数 $\hat{m}(k;x,t)$ が求まった。式 (11) を式 (5) に代入すると

$$J_{m_k=\hat{m}_k} = \iint_D f(x,t)\log\frac{f(x,t)}{\sum_{\forall k} q_k(x,t;\Theta)}dxdt \qquad (12)$$

となり、結局、以上のクラスタリングは、観測分布 $f(x,t)$ と k についてすべてのモデル関数 $q_k(x,t;\Theta)$ を重畳した分布とのカルバック-ライブラー (KL) 情報量 (擬距離) を最小化することと数学的に等価であることが分かる。また、式 (11) の導出結果は、k を内部状態変数 (あるいは隠れデータ) と見なし、$q_k(x,t;\Theta)$ を完全データの確率密度 $p(k,x,t|\Theta)$ として解釈すると、確率則を一切用いなくても EM(Expectation-Maximization) アルゴリズムの収束性が証明できることを示しており、興味深い。EM アルゴリズムとの対応関係は、式 (6) と式 (11) を Q 関数

$$Q(\Theta,\tilde{\Theta}) = \sum_{\forall k}\iint_D \overbrace{p(k|x,t,\Theta)}^{\text{隠れデータ事後確率}}\ \overbrace{f(x,t)}^{\text{観測確率密度}}\log\overbrace{p(k,x,t|\tilde{\Theta})}^{\text{完全データ尤度}}dxdt$$

$$\left(p(k|x,t,\Theta) = \frac{p(k,x,t|\Theta)}{p(x,t|\Theta)} = \frac{p(k,x,t|\Theta)}{\sum_{\forall k} p(k,x,t|\Theta)}\right) \qquad (13)$$

と見比べるとより明白になる。ただし、ここでは $k,x,t \in \Omega$ は確率変数であり、

$$\iint_D f(x,t)dxdt = 1,\quad \sum_{\forall k}\iint_D p(k,x,t|\Theta)dxdt = 1$$

を満たす。さて、一方で Θ は、エネルギー分配関数 $m(k;x,t)$ を固定として、

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, J = \underset{\Theta}{\operatorname{argmax}}\, I(\Theta) = \underset{\tilde{\Theta}}{\operatorname{argmax}}\, Q(\Theta,\tilde{\Theta}) \qquad (14)$$

で最適化すれば良い。この更新式はモデル関数 $q_k(x,t;\Theta)$ の定義次第で決まることになる。以上の多重音解析一般解法をスペクトル時間構造化クラスタリング (Spectro-Temporal structured clustering; STC) と呼ぶ。

3. モデルの定式化

3.1 ガウス基底調波時間構造化モデル

ここでは、時間周波数エネルギー分布関数モデル $q_k(x,t;\Theta)$ を定式化する。今、目的は音楽の音響特徴量抽出なので、簡単のため調和性をもつ音響信号だけを対象としよう。非調和性のモデリングについては今後検討することにする。STC の枠組の中で、調和性をもつ音響信号に対象を限定したものを調波時間構造化クラスタリング (Harmonic-Temporal structured Clustering; HTC) と呼び、$q_k(x,t;\Theta)$ を調波時間構造化モデル (HTM) と呼ぶ。

まず、k 番目の音響イベントが発音開始してから減衰するまでの間の対数基本周波数の時間軌跡を多項式

$$\mu_k(t) = \mu_{k0} + \mu_{k1}t + \mu_{k2}t^2 + \cdots \qquad (15)$$

で表すことにすると、時刻 t における $q_k(x,t;\Theta)$ の切口は、図 3 のような対数基本周波数 $\mu_k(t)$ の調波構造をなすはずである。また、n 次の対数倍音周波数は $\mu_k(t) + \log n$ である。ガボールウェーブレット変換では、ガウス窓が周波数領域で畳み込まれることになるので、周波数成分の拡散エネルギー分布はガウス分布関数で十分良く近似できると考えられる。次に、各周波数成分パワーは時間に伴って一般に連続的に変化していくものなので、そのパワーエンベロープを $u_{kn}(t)$ とする。ただし、式 (4) を満たす必要があるので $u_{kn}(t)$ は正規化可能、すなわち無限区間積分が有界な関数であることを前提とし、

$$\forall k, \forall n, \int_{-\infty}^{\infty} u_{kn}(t)dt = 1, \qquad (16)$$

を条件とする関数とする。以上より、n 次倍音成分のエネルギー分布はガウス分布関数とパワーエンベロープ関数 $u_{kn}(t)$ と倍音エネルギー比の積

$$\nu_{kn}u_{kn}(t) \times \underbrace{\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2}}}_{x=\mu_k(t)+\log n\ \text{を中心とするガウス分布}} \qquad (17)$$

で表される。σ_k は分布の拡散パラメータ、ν_{kn} は

$$\forall k, \sum_{\forall n}\nu_{kn} = 1 \qquad (18)$$

を満たす、他の倍音成分との間の相対的な強度を表す重みパラメータであり、音色を決定づける要素の一つである。以上より、k 番目の HTM $q_k(x,t;\Theta)$(図 2) は、N 個の倍音成分時間パターンをすべて重畳加算した

$$q_k(x,t;\Theta) = w_k\sum_{n=1}^{N}\frac{\nu_{kn}u_{kn}(t)}{\sqrt{2\pi}\sigma_k}e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2}} \qquad (19)$$

と表される。w_k は、k 番目の音響イベントのエネルギーを表す。さらに、K 個の HTM を重畳加算した

$$L(x,t;\Theta) = \sum_{k=1}^{K} q_k(x,t;\Theta)\quad (k=1,\cdots,K) \qquad (20)$$

は、時間周波数平面に拡散したエネルギーの観測パターン全体に対する関数モデルになる。

$u_{kn}(t)$ の関数モデリングは本研究の中でも重要な要素に位置づけられる。どの楽器が用いられても安定的に動作する汎用な音響特徴量抽出手法が望まれるが、このような設定問題は典型的な暗中 (ブラインド) 問題である。その場合、特定の楽器

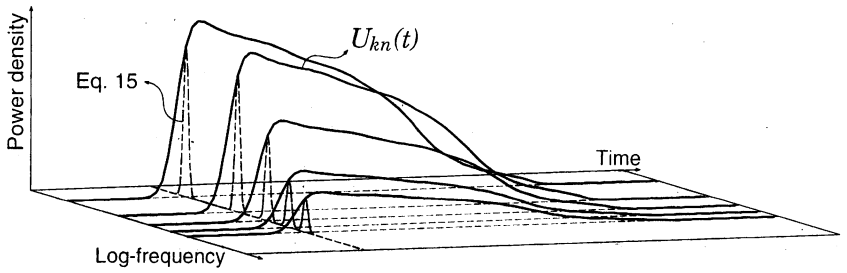


図 2 k-th harmonic-temporal-structured model (HTM) $q_k(x,t;\Theta)$ (Eq. 19)

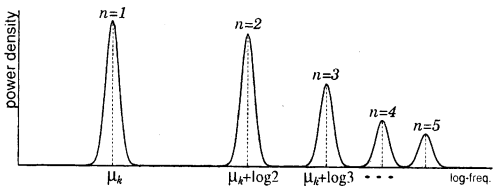


図 3 Cutting plane of $q_k(x,t;\Theta)$ at time t (Eq. 17)

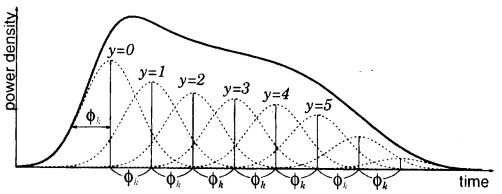


図 4 Power envelope function $U_{kn}(t)$ (Eq. 21)

表 1 HTM パラメータの音響特徴量としての意味

記号	意味
$\mu_k(t)$	音響イベント中の音高軌跡の多項式 (0 次の多項式の場合は時間軸に平行な軌跡)
w_k	音響イベントのエネルギー
v_{kn}	n 番目の倍音成分の相対エネルギー ($\{v_{kn}\|n=1,\cdots,N\}$ は音色特徴量の要素になりうる)
$u_{knγ}$	n 番目の倍音成分のパワーエンベロープの概形
τ_k	音響イベントのオンセット時刻
$\gamma\phi_k$	音響イベントの継続時間

の物理的な発音メカニズム (駆動特性と共振特性) に特化して $U_{kn}(t)$ をモデリングするより、むしろ幅広く柔軟に対応できるモデリングの検討が重要であると我々は考える。

そのためには、$U_{kn}(t)$ は、連続的であり、すべての t において非負であり、$t \to \infty$, $t \to -\infty$ で 0 に収束し、時間方向に伸縮自在であり、いかなる曲線にも良くフィットするような関数であるのが望ましい。また、式 (14) と式 (16) を達成できるにはあらゆる t で微分可能であり、無限区間積分が計算可能であるべきである。そこで、我々は

$$U_{kn}(t) = \sum_{\gamma=0}^{\Upsilon-1} \frac{u_{kn\gamma}}{\sqrt{2\pi}\phi_{kn}} \exp\left(-\frac{(t-\tau_k-\gamma\phi_{kn})^2}{2{\phi_{kn}}^2}\right) \quad (21)$$

で与えられる特殊な拘束つき混合ガウス分布関数モデルを提案する。ただし、τ_k は先頭のガウス分布の中心で、発音開始時刻の推定値に相当する。また $u_{kn\gamma}$ は各ガウス分布にかかる重み係数であり、

$$\forall k, \forall n, \sum_{\forall\gamma} u_{kn\gamma} = 1 \quad (22)$$

を満たすこととする。この係数を自由に定めることで曲線をさまざまに変形することができる。この関数は、ガウス分布を基底に選んでいることからウェーブレット変換における時間方向のスペクトル漏洩を良く表現したモデルになっている。(一般にガボールウェーブレット変換では、ある時間長でしか発音されていなくても、ガウス窓の影響でエネルギーがその範囲外にも拡散される。) この関数における独自性は、Υ 個のガウス分布が共通の標準偏差値 ϕ_{kn} をもち、その値と等しい間隔で常に配置されるという拘束条件をもつ点にあり、この拘束は各ガウス成分を孤立させないようにするだけでなく、ϕ_{kn} の大きさに依存して関数全体が時間方向に線形伸縮する性質を与える。

以上で与えられる HTM のパラメータはすべて、音楽情報検索目的において極めて有用な特徴量になりうる。各パラメータ特徴量をまとめて表 3.1 に示す。

3.2 カーネル・サブクラスタリング

HTM は基底関数の線形結合で構成されていることから、各クラスタをさらに {n, γ} ラベルつきのサブクラスタに分解することで、式 (14) の解析的最適解の計算が可能になる。

$q_k(x,t;\Theta)$ は {k, n, γ} ラベルつきのカーネル密度関数 $S_{knγ}(x,t;\Theta)$ の和の形

$$q_k(x,t;\Theta) = \sum_{\forall n}\sum_{\forall\gamma} \underbrace{w_k H_{kn}(x,t) E_{kn\gamma}(t)}_{\equiv S_{kn\gamma}(x,t;\Theta)} \quad (23)$$

$$\begin{cases} H_{kn}(x,t) & \equiv \frac{v_{kn}}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-u_k(t)-\log n)^2}{2\sigma_k^2}} \\ E_{kn\gamma}(t) & \equiv \frac{u_{kn\gamma}}{\sqrt{2\pi}\phi_{kn}} e^{-\frac{(t-\tau_k-\gamma\phi_{kn})^2}{2\phi_{kn}^2}} \end{cases}$$

に書き直すことができる。ここで、新たにクラスタ分配関数 $m(n,\gamma;k,x,t)$ を導入する。$m(n,\gamma;k,x,t)$ は、

$$\forall k, \sum_{\forall n,\forall\gamma} m(n,\gamma;k,x,t) = 1, \quad 0 \le m(n,\gamma;k,x,t) \le 1,$$

を満たす、k 番目のクラスタ $m(k;x,t)f(x,t)$ を {n, γ} ラベルつきのサブクラスタに分配するための関数である。以上の条件を満たすいかなる $m(n,\gamma;k,x,t)$ に関して以下の不等式

$$J_k \equiv \iint_D m(k;x,t)f(x,t)\log\frac{m(k;x,t)f(x,t)}{\sum_{\forall n,\forall y} S_{knu}(x,t;\Theta)}dxdt$$

$$\leqq \tilde{J}_k \equiv \sum_{\forall n,\forall y}\iint_D m(k;x,t)m(n,y;k,x,t)f(x,t)$$
$$\log\frac{m(k;x,t)m(n,y;k,x,t)f(x,t)}{S_{knu}(x,t;\Theta)}dxdt \quad (24)$$

が成り立ち、等号は

$$m(n,y;k,x,t) = \frac{S_{knu}(x,t;\Theta)}{\sum_{\forall n}\sum_{\forall y} S_{knu}(x,t;\Theta)} \quad (25)$$

のときに成立する。以上の証明は、2. と同様の導出過程で示すことができるのでここでは省略する。クラスタ分配関数が式 (25) のとき $\tilde{J}=\sum_{\forall k}\tilde{J}_k$ はもとの目的関数 $J=\sum_{\forall k}J_k$ と等しくなり、式 (14) を達成するには $\tilde{J}$ を最小化すれば良いことになる。以上より、パラメータ更新式は

$$\hat{\Theta} = \underset{\Theta}{\mathrm{argmax}}\sum_{\forall k,\forall n,\forall y}\iint_D \overbrace{m(k;x,t)m(n,y;k,x,t)}^{\equiv m(k,n,y;x,t)} f(x,t)$$
$$\log S_{knu}(x,t;\Theta)dxdt \quad (26)$$

より導出すれば良い。

4. MAP 推定としての解釈

4.1 事前分布の仮定

正規化したエネルギー密度分布 $f(x,t)$ を確率密度関数、$L(x,t;\Theta)=\sum_{\forall k}q_k(x,t;\Theta)$ をパラメータの条件つき確率密度関数 (モデル尤度関数) と捉えると、本問題は対数尤度の期待値の最大化 (最尤推定)

$$\hat{\Theta}_{ML} = \underset{\Theta}{\mathrm{argmin}}\, J_{m_k=\nu_k}$$
$$\Leftrightarrow \underset{\Theta}{\mathrm{argmax}}\left\langle \log L(x,t;\Theta)\right\rangle_{f(x,t)} \quad (27)$$

と同じ形をとる。ただし、$\langle\cdot\rangle_\Omega$ は期待値をさす。このような観点に立てば、パラメータに事前分布を仮定し、最大事後確率 (Maximum *A Posteriori*) 推定に拡張できることに自然に気づく。MAP 推定は、パラメータに関する経験的な統計を含めた形で、対数事後確率の期待値

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\mathrm{argmax}}\left\langle \log L(x,t;\Theta)+\log p(\Theta)\right\rangle_{f(x,t)} \quad (28)$$

を求める問題になる。事前分布 $p(\Theta)$ は、パラメータが「常識的」な範囲を逸脱しないように制約を与えるためのペナルティ関数のように働く。観測に関係ないこの事前分布項と、観測によって決まる対数尤度項との二つの間の最も良い妥協点を探すことが MAP 推定の直感的な理解である。この働きからは、本問題における次のような 2 つの効果が期待される。

まず、真の対数基本周波数軌跡を $F(t)$ とすると、例えば $\mu_k(t)=F(t)-\log r, (r\in\mathbb{N}, r>2)$ のとき、$r\times n$ 次の ν_{kn} がすべて 0 であるモデルは、n が十分大きければ $\mu_k(t)=F(t)$ のときのモデルとまったく同等である。つまり、ν_{kn} が完全に自由だと、$\mu_k(t)=F(t)-\log r$ のモデルは $\mu_k(t)=F(t)$ のモデルを完全に包含する。従ってこの場合、分布間距離を目的関数とする本問題では、$\mu_k(t)$ が $F(t)/r$, $r\in\mathbb{N}$ すべてにおいて大域最小点を与える。しかしながら、$r\times n$ 次以外の周波数成分がすべて 0 であるような調波構造は極めて稀れで、通常は経験的な意味で「不自然」と捉えることができる。従って、我々のもつ調波構造に対する常識的な規範を ν_{kn} の拘束条件として目的関数に加えることで、$\mu_k(t)$ が $F(t)$ のときと $F(t)/r$, $r\in\mathbf{N}$ のときとを差別化することができ、基本周波数推定の精度向上に寄与できる。

一方、式 (21) で与えられるパワーエンベロープ関数の中の係数 u_{knu} に完全に自由度を与えると、一つの HTM だけで、同一音高が続けて複数回発音された場合の一連のパワースペクトル時系列を表現できてしまうことがある。本来このような状況のときは、各発音ごとのオンセット時刻を推定したいはずであるが、この場合、複数の谷間があるのこぎり波のような「不自然」なパワーエンベロープをもつ単一の音響信号として解釈される。従って、前述の例同様、ある程度の常識的な規範を u_{knu} の拘束条件として目的関数に加えることで上のような極端な推定値から回避できるようになり、オンセット時刻推定の精度向上に寄与できる。

ここでは後藤が提案した事前分布 [18]

$$\begin{cases} p(\nu_k) \equiv \dfrac{1}{Z_\nu}\exp\left(-d_\nu\sum_{\forall n}\bar{\nu}_n\log\dfrac{\bar{\nu}_n}{\nu_{kn}}\right) \\ p(u_{kn}) \equiv \dfrac{1}{Z_u}\exp\left(-d_u\sum_{\forall y}\bar{u}_y\log\dfrac{\bar{u}_y}{u_{kny}}\right) \end{cases} \quad (29)$$

$$\sum_{\forall n}\bar{\nu}_n = 1, \quad \sum_{\forall y}\bar{u}_y = 1 \quad (30)$$

を用いる。この事前分布は負の KL 情報量の指数で定義され、ν_{kn} および u_{kny} が $\bar{\nu}_n$、$\bar{u}_y$ と完全に一致するときに最大値をとる。ただし、d_r、d_c は事前分布の寄与の大きさを表し (小さいほどペナルティの効力は下がり、大きいほど上がる)、Z_r、Z_c は正規化係数である。なお、この事前分布は、EM アルゴリズムの M ステップにおける ν_{kn} および u_{kny} の更新式の計算を大幅に簡単化する利点がある。ただし、同様の手段としてディリクレ分布も事前分布として適用できる。

w_k、ν_{kn}、u_{kny} に関する未定乗数 γ_w、$\gamma_\nu^{(k)}$、$\gamma_u^{(kn)}$ によるラグランジュ項と事前分布項を含めると、M ステップでは結局

$$\hat{\Theta}_{MAP} = \underset{\forall k}{\mathrm{argmax}}\sum\Bigg(\Bigg(\sum_{\forall n}\sum_{\forall y}\iint_D m(k,n,y;x,t)$$
$$f(x,t)\log S_{knu}(x,t;\Theta)dxdt\Bigg)$$
$$-d_\nu\sum_{\forall n}\bar{\nu}_n\log\frac{\bar{\nu}_n}{\nu_{kn}} - d_u\sum_{\forall n}\sum_{\forall y}\bar{u}_y\log\frac{\bar{u}_y}{u_{kny}}$$
$$-\gamma_\nu^{(k)}\Bigg(\sum_{\forall n}\nu_{kn}-1\Bigg) - \sum_{\forall n}\gamma_u^{(kn)}\Bigg(\sum_{\forall y}u_{kny}-1\Bigg)\Bigg)$$
$$-\gamma_w\Bigg(\sum_{\forall k}w_k-1\Bigg) \quad (31)$$

を解けば良いことになる。

5. パラメータ更新式

抽出する特徴量の次元を減らす目的で、対数基本周波数の時間軌跡は時間軸と平行 (0 次多項式) と仮定する。すなわち、$\mu_k(t) \approx \mu_{k0}$ と簡単化する。また、一つの HTM の中の各倍音成分のパワーエンベロープは互いに相似であると仮定する。すなわち、各 HTM のパワーエンベロープは $\nu_{kn}u_{ky}(t)(u_{ky(t)}$ は n に依らない) で表す。今回の我々の目標は、精密な演奏解析を主眼としたものではなく、むしろ大局的な演奏情報を抽出することにあるのでこれらの仮定は必ずしも致命的なものではない。ただし提案手法が前者のような目的のための手段としても応用できる枠組であることは留意されたい。さて、

$$\ell_{kny}(x,t) \equiv \hat{m}(k;x,t)\hat{m}(n,y;k,x,t)f(x,t)$$

とおくと (23) より、式 (31) を解くと、

$$w_k^{(i+1)} = \sum_{\forall n,\forall y}\iint_D \ell_{kny}(x,t)dxdt$$

$$\mu_{k0}^{(i+1)} = \frac{\sum_{\forall n,\forall y}\iint_D (x-\log n)\ell_{kny}(x,t)dxdt}{w_k^{(i+1)}}$$

$$\tau_k^{(i+1)} = \frac{\sum_{\forall n,\forall y}\iint_D (t-y\phi_k^{(i)})\ell_{kny}(x,t)dxdt}{w_k^{(i+1)}}$$

$$\nu_{kn}^{(i+1)} = \frac{d_\nu\bar{\nu}_n + \sum_{\forall y}\iint_D \ell_{kny}(x,t)dxdt}{d_\nu + w_k^{(i+1)}}$$

$$u_{ky}^{(i+1)} = \frac{d_u\bar{u}_n + \sum_{\forall n}\iint_D \ell_{kny}(x,t)dxdt}{d_u + w_k^{(i+1)}}$$

$$\phi_k^{(i+1)} = \frac{-\Lambda_k + \left(\Lambda_k^2 + 4\sum_{\forall y}\int \Upsilon_{ky}(t)^2(t-\tau_k)^2dt\right)^{1/2}}{2w_k^{(i+1)}}$$

$$\left(\begin{array}{l}\Lambda_k = \sum_{\forall n,\forall y}\iint_D y(t-\tau_k)\ell_{kny}(x,t)dxdt \\ \Upsilon_{ky}(t) = \sum_{\forall n}\int \ell_{kny}(x,t)dx\end{array}\right)$$

$$\sigma_k^{(i+1)} = \left(\frac{\sum_{\forall n,\forall y}\iint_D (x-\mu_{k0}^{(i)}-\log n)^2\ell_{kny}(x,t)dxdt}{w_k^{(i+1)}}\right)^{1/2}$$

のように更新式の解析解が得られる。

6. 評価実験

6.1 条　　件

提案手法の音楽音響特徴量抽出としての性能を確認するため、RWC 研究用音楽データベース [21] の実音楽音響信号を対象として評価実験を行った。用いた実験データを表 2 に記す。

f(x,t) は、ガボールウェーブレット変換 (サンプリング周波数 16kHz、時間分解能 16ms、最低周波数 60Hz、周波数分解能 12cent) により解析した。解析する時間周波数平面の時間区間は連続する 80 フレーム (1.28s) とした。HTM の対数基本周波数およびオンセット時刻 ($\mu_{k0},\tau_k|k=1,\cdots,K$) のパラメータ初期値は、f(x,t) の極大点のうちエネルギーの大きいものから 20 点抽出して、それらの時間周波数平面上での座標とした。また、音響イベントの推定総数は w_k が閾値より大きい HTM の個数とした。実験条件をまとめて 4 に記す。

表 4　Experimental Conditions

frequency analysis	Sampling rate	16 kHz
	frame shift	16 ms
	frequency resolution	12 0 cent
	frequency range	60–3000 Hz
HTC	initial # of HTMs	20
	# of partials: N	6
	# of kernels in $U_k(t)$: Y	10
	$\bar{\nu}_n$	$0.6547 \times n^{-2}$
	$\bar{u}_y$	$0.2096 \times e^{-0.2y}$
	d_ν, d_u	0.04
	range of analyzing segment	80 frames (1.28 s)
	# of analyzing segments	21 (total time: 24 s)
PreFEst [18]	pitch resolution	20 cent
	# of partials	8
	# of tone models	200
	standard deviation of Gaussian	3.0
	$\bar{\nu}_n$	$0.6547 \times n^{-2}$
	d (prior contribution factor)	3.0

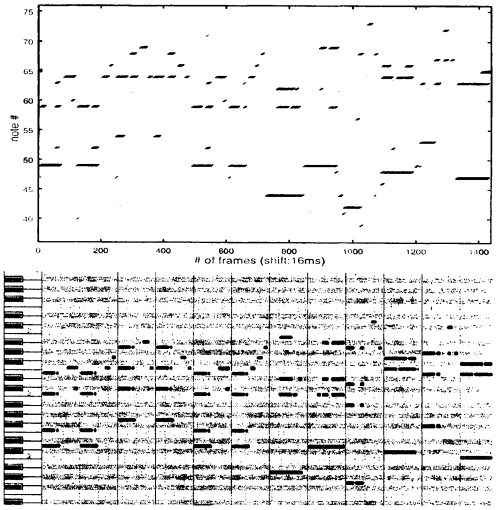


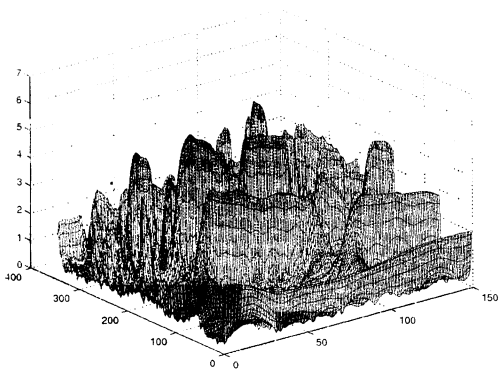
図 6　$\mu_{k0},\tau_k,\Upsilon\phi_k$ の推定結果 (上) と付属参照用 MIDI データ (下) のピアノロール形式表示。

音高の正解精度は、DP(Dynamic Programming) に基づいた自動計算法を実装した。実装手順は紙面の都合上省略する。この自動計算では、置換誤りを脱落と挿入の二重の誤りと判断するため、場合によっては正解率が負となることがある。

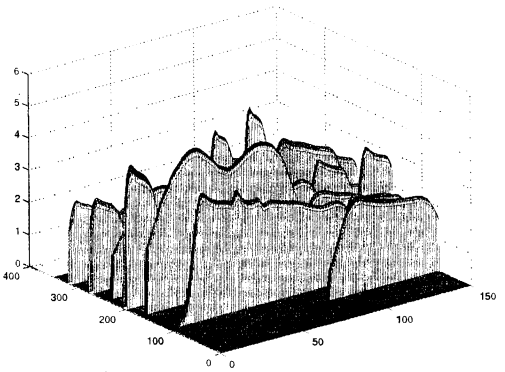
6.2 結　　果

提案手法の比較対象として 'PreFEst' [18] を選んだ。PreFEst は front-end 部、core 部、back-end 部の 3 段階の処理で構成されているが、今回は core 部だけを実装した (これを PreFEst-core と以後呼ぶ)。PreFEst-core はフレームごとに音高尤度を出力する処理であり、音源数を推定する具体的な手続きはないため、提案法と同様に閾値により音高候補のトランケーションを行った (尤度が閾値以上となる基本周波数を、最も近い音階に量子化したものを推定音高とした)。

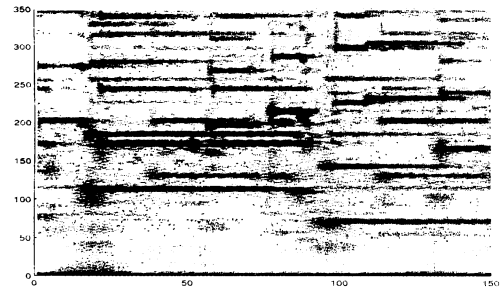
図 6 に、データ (1) に対し提案法により求まった対数基本周波数 (時間軌跡) とオンセット時刻と音長を、対応する付属参照



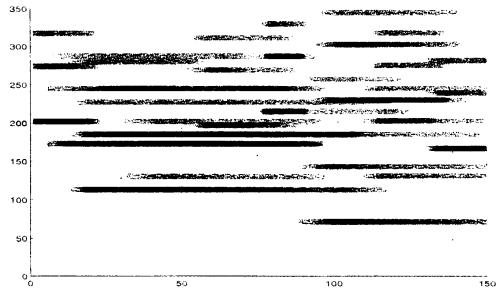
(a) f(x, t) の 3D 表示



(b) (a) に対して求めた HTM 重畳分布



(c) f(x, t) のスペクトログラム



(d) (b) の濃淡表示

図 5　f(x, t) と求めた HTM 重畳分布の 3D および濃淡表示。

表 2　RWC 研究用音楽データベース [21] より抜粋した実験データ

Symbol	Title (Genre)	Catalog number	Composer/Player	Instruments	# of frames
data(1)	Crescent Serenade (Jazz)	RWC-MDB-J-2001 No. 9	S. Yamamoto	Guitar	4427
data(2)	For Two (Jazz)	RWC-MDB-J-2001 No. 7	H. Chubachi	Guitar	6555
data(3)	Jive (Jazz)	RWC-MDB-J-2001 No. 1	M. Nakamura	Piano	5179
data(4)	Lounge Away (Jazz)	RWC-MDB-J-2001 No. 8	S. Yamamoto	Guitar	9583
data(5)	For Two (Jazz)	RWC-MDB-J-2001 No. 2	M. Nakamura	Piano	9091
data(6)	Jive (Jazz)	RWC-MDB-J-2001 No. 6	H. Chubachi	Guitar	3690
data(7)	Three Gimnopedies no. 1 (Classic)	RWC-MDB-C-2001 No. 35	E. Satie	Piano	6571
data(8)	Nocturne no.2. op.9-2(Classic)	RWC-MDB-C-2001 No. 30	F. F. Chopin	Piano	7258

用 MIDI データとともにピアノロール形式で示す。また、観測分布と求めた重畳モデル分布の 3D および濃淡表示したものを図 5 に示す。

閾値による音高候補のトランケーションでは、閾値の大きさに応じた挿入誤りと脱落誤りの数との間にはトレードオフがあるが、さまざまな閾値を試した中で最も良い音高正解率を比較することで、両手法の潜在能力の限界を知ることができるはずである。PreFEst-core と HTC の各データに対するそれぞれの設定閾値における音高正解率の結果を表 3 に示す。各データで試した閾値のうち最も高かった正解率は太字で表記してある。両者を比較すると、HTC がいずれのデータに対しても上回っており、フレームごとの独立なスペクトルモデリングよりもスペクトルの時間周波数構造を同時にモデリングしたことの有効性を示すことができた。

7. 結　　論

本報告で、音楽検索のフロントエンドである音響特徴量抽出技術の開発を目標として、パワースペクトルの時間周波数構造の 2 次元幾何モデリングに基づきエネルギーを音響イベントごとにクラスタ化する方法論、HTC を提案した。我々はこの問題が、観測分布とモデル分布との分布間距離最小化と等価であることを示し、さらに MAP 推定と同形問題として拡張できることを示した。また、サブクラスタリングにより M ステップ更新式の解析解の導出を可能にした。

本報告で述べた本手法にはいまだいくつも興味深い考察事項が残されている。音響イベントの総数の自動推定の手段、対数基本周波数軌跡のより洗練したモデリング、非調和性を考慮したモデリングなどを今後は取り組みたい。また本手法のアプリ

表 3 PreFEst-core [18] と HTC の性能比較．(A)-(H) と (I)-(P) の列は異なるトランケーション周波数に対する正解精度を示す．各周波数は以下とした．(A)2.0×10^8, (B)2.5×10^8, (C)5.0×10^8, (D)7.5×10^8, (E)1.0×10^9, (F)1.5×10^9, (G)$1.7.5 \times 10^8$, (H)2.0×10^8, (I)2.5×10^8, (J)$2.7.5 \times 10^8$, (K)7.5×10^9, (L)1.0×10^{10}, (M)2.0×10^{10}, (N)3.0×10^{10}, (O)4.0×10^{10}, (P)5.0×10^{10}, (Q)6.0×10^{10}, (R)7.0×10^{10}

	Accuracy(%)																	
	従来法 (PreFEst-core, [18])										提案法 (HTC)							
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)	(P)	(Q)	(R)
data(1)	56.6	62.49	75.9	81.6	83.3	**84.6**	83.0	81.5	78.4	75.8	69.5	74.8	83.9	84.8	88.2	**88.8**	88.7	85.1
data(2)	68.7	**69.6**	66.3	59.0	53.7	36.3	32.4	30.3	26.8	26.5	84.3	88.2	**90.6**	82.5	75.7	72.3	67.9	61.9
data(3)	-20.8	-7.3	31.7	47.8	56.9	65.1	69.5	71.9	**75.5**	71.8	68.8	70.0	77.6	80.0	**80.2**	77.4	73.3	73.4
data(4)	55.1	56.8	60.7	63.3	63.1	63.6	**64.1**	62.3	60.6	60.2	82.6	83.0	**83.8**	82.4	82.8	82.0	81.5	76.5
data(5)	50.7	53.2	**61.0**	60.0	58.8	59.3	57.6	58.0	57.5	49.7	76.3	79.3	79.4	**81.7**	77.6	76.2	76.5	72.8
data(6)	-7.2	6.6	37.9	51.1	57.7	65.9	65.6	**66.7**	66.3	65.7	77.5	79.6	81.7	82.7	**84.4**	82.3	81.4	80.7
data(7)	51.6	54.1	**62.7**	52.4	47.0	45.9	42.7	41.1	42.2	42.7	**72.1**	69.9	70.3	68.3	66.9	63.1	61.5	62.0
data(8)	20.8	22.9	36.6	**42.5**	38.5	39.1	38.8	37.7	32.7	30.6	73.7	**75.9**	75.6	72.2	67.6	61.1	48.9	46.7

ケーションについても検討していきたい。

文　献

[1] 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.

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

[3] 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, 1988.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] K. Kashino, H. Murase, "A Music Stream Segregation System based on Adaptive Multi-agents," In Proc. *IJCAI-97*, pp. 1126–1131, 1997.

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

[16] 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.

[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] M. Goto, "A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals," *ISCA Journal*, Vol. 43, No. 4, pp. 311–329, 2004.

[19] 高橋佳吾, 西本卓也, 嵯峨山茂樹, "対数周波数逆畳み込みによる多重音の基本周波数解析," 情報処理学会研究報告, 2003-MUS-53-13, pp. 61–66. 2003.

[20] 亀岡弘和, 齊藤翔一郎, 西本卓也, 嵯峨山茂樹, "Specmurt における準最適共通調波構造パターンの反復推定による多声音楽信号の音高可視化と MIDI 変換," 情報処理学会研究会報告, 2004-MUS-56-7, pp. 41–48. 2004.

[21] 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.