2007－MUS－71 (18)
2007／8／2

パート除去を目的とした楽譜と音響信号のアラインメント手法の検討

松本恭輔† 西本卓也†
小野順貴† 嵯峨山茂樹†

本稿では、楽譜情報を利用して音楽音響信号から1パート除去等の加工を目的とした、楽譜と音響信号の詳細なアラインメントについて議論する。実演奏は厳密に楽譜通りではなく、テンポ変動の他に各音のデビエーションを含む。そのため、上述の加工に向けて、一音毎の発音時刻・音長等の詳細な情報を求める必要があるが、テンポ推定のみ行う従来のアラインメント手法では、この問題への対応は困難である。これに対して我々は、テンポ変動、各音のデビエーション、音色の変動を考慮した、「楽譜からの音楽音響信号生成モデル」に基づき、事後確率最大化の枠組で詳細なアラインメントをとるアプローチを提案する。この枠組に従う一手法の検討として、実験を行い、確かに微小変動を扱える枠組であること、デビエーションのモデルにさらに検討の余地があることを定性的に確認した。

Alignment of Music Score to Audio for Part Removal from Ensemble Music

Kyosuke Matsumoto,† Takuya Nishimoto,† Nobutaka Ono† and Shigeki Sagayama†

This paper describes an approach to alignment of music score to audio which does not estimates only tempo but also deviations for each note in real performance. Estimating the deviations is required for score-guided 1 instrumental part removal or any other manipulation of musical audio signal. For this problem, existing alignment algorithms are not sufficient because they deal with only tempo. On the other hand, our approach models how musical signal is generated from score, considering (i) tempo variation, (ii) deviations in each note and (iii) variation of acoustic signal features, and then, estimates tempo and deviations with maximum a posteriori probability based on the model. Simulation quantitably show our approach has advantage of processing deviation and needs more consideration especially on deviation modeling.

1. はじめに

本稿では、音楽音響信号上の1パートの音を除去することを目的とした、楽譜と音響信号の詳細なアラインメントについて論じる。これは、楽譜の各音符に対応する音の時間区間・音高等を詳細に求めようとするものであり、「除去」以外の応用も見込める。例えば、音楽演奏解析の研究では、各音のデビエーションが重要な意味を持つ。これらの情報が自動で精度よく求められれば有用である。また、音楽音響信号の加工を目的とした研究には、アラインメントの取れた楽譜を前提としているものがあるが[1]、アラインメントの取れていない楽譜からの加工が可能になれば、より使いやすく、幅広い楽曲に適用可能なシステムとなるだろう。詳細なアラインメント技術は上記のように、楽譜が既知である場合に音楽音響信号の詳細な解析を行う手法として重要である。

楽譜と音響信号のアラインメントの従来研究は、録音へのランダムアクセス、自動伴奏システム等への応用を想定し、音響信号上の時刻と楽譜上の拍を対応付ける問題を扱ってきた。そのため、従来手法の多くは、Orioらによる学習無しのオフライン手法[2]や隠れマルコフモデルを用いた学習有りのオンライン手法[3]そして、Raphaelら[4]によるテンポの微小変動を導入したグラフィカルモデルに基づく手法に代表されるように、時間フレーム毎の動的計画法を基礎にしたものであった。これらの従来手法はいずれもテンポ推定のみを行う手法であると言える。

一方我々が想定する応用は、楽譜情報を利用した音響信号の加工である。実演奏にはテンポ変動のみならず、各音の音長・発音時刻・音高の微小変動等様々な変動が含

† 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

まれる。そのため、ある特定の音を適切に除去するには、一音毎に詳細なアラインメントをとる必要が有る。例えば、図 1 に示すように、楽譜上で同一拍に属する複数音がずれて発音される場合や、楽譜上の音価より実際の音が長い場合には、テンポ推定のみを行う従来手法での詳細なアラインメントは不可能である。実際に我々は、テンポ推定を行う従来法[2] と、対象楽器を限定した上で、ヒューリスティックであるものの各音の発音・消音時刻を推定する手法[5] のそれぞれを用いて 1 パート除去を行い、後者がより適切な除去を達成することを確かめている (図 2 参照)。

そこで我々は、「楽譜からの音楽音響信号生成モデル」に基づく詳細なアラインメントへのアプローチを提案する。これは、(1) 従来法で扱われてきたテンポ変動、(2) 詳細なアラインメントに必要不可欠な各音のデビエーション、そして (3) 音色変動が、楽譜に加わることで実演奏の音響信号が生成される、というモデルに基づいて、事後確率最大化 (MAP) 推定により、詳細なアラインメントを取るものである。このアプローチの一番の狙いは、各音のデビエーションをモデル化して、詳細なアラインメントを達成することにあるが、さらに音色変動も考慮することで、楽譜と音響信号の間の適応的距離尺度を利用したロバストな推定、アラインメントと同時に音色情報を取得可能な演奏情報解析手法への発展等が期待できる。

本稿の構成は、以下の通りである。まず、 第 2 節 で、除去を目的としたアラインメント問題について議論する。次に 第 3 節 で、楽譜からの音楽音響信号の生成過程をモデル化する。第 4 節 で、確率的逆問題としてのアラインメント問題を定式化し、解法を示す。提案手法の適用例を 第 5 節 で示し、 第 6 節 でまとめと展望を述べる。

2. 問題設定

2.1 対象とする音響信号の取扱い

楽譜とのアラインメントの難易は楽器に依存すると考えられる。リズムを刻むような打楽器が含まれる場合は、それに基づいてテンポの追跡が行えるであろう。一方、調波性の楽器のみで構成されるクラシックの楽曲の演奏では、しばしば楽譜からの大きな変動が見られ、それが演奏解析の対象となることが多い。そこで、本研究では、ドラム等の非調波性の強い楽器を含まない楽曲を対象にしたアラインメントの問題を考える。

問題の音響信号入力情報としては短時間パワースペクトルを用いることにする。音楽音響信号は、残響を多く含んでいたり、ライン入力からのミキシング処理で作られたりして、位相情報は必ずしも有効利用できない。音楽音響信号から指定された音を除去する問題においても、

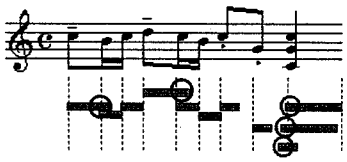

図 1 発音時刻・音長の変動が有る場合、推定されたテンポ (点線の区間) に従って除去しても，消し残しが生じる。

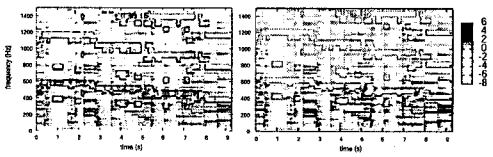

図 2 除去処理後スペクトログラム (除去するのは線で囲まれた部分): 従来アラインメント[2](左) で除去できなかった部分が発音時刻・音長推定[5](右) を利用し除去できる。(音源は RWC 研究用音楽データベース (ジャズ) から RWC-MDB-J-2001 No.11 より)

本来ならば波形領域での減算処理を行うべきだが、同様の理由で困難であり、パワースペクトル領域での操作を考える方が妥当であろう。

2.2 楽譜情報の取扱い

本研究で扱う楽譜情報は、基本的かつ曖昧性無く工学的に扱いやすい「各音の音名・発音拍・音価と、楽譜に明示されたテンポ」に限定する。楽譜情報の表現にはさまざまな形式があるが、多くの場合は容易に MIDI データに変換できるので、本問題での具体的な入力形式には、「アラインメントの取られていない」MIDI ファイル (Standard Midi File; SMF) 等が適当であろう。MIDI には、楽器情報を含むものが存在するので、将来的には、各パートの音色を利用した、アラインメント問題も考えられる。

2.3 何を推定するか

適切な除去を行うためには、音響信号上の各除去対象音に関して「発音時刻・音長・音高」が必要である。また、これらの情報に対してテンポは独立な情報ではなく、同時に推定が行えるはずである。そこで、「各音の発音時刻・音長・音高、演奏のテンポ」を推定する問題を扱う。ただし、音名に対して一定した音高が対応するピアノ曲など音高推定が不要な場合もある。

3. 楽譜からの音楽音響信号生成モデル

3.1 楽譜からの音楽音響信号の生成過程

楽譜と音響信号は、同一の楽曲に関する情報であるが、

(1) テンポ変動
(2) 各音の発音時刻・音長・音高のデビエーション (以下、単に「デビエーション」と呼ぶ)
(3) 音色の変動

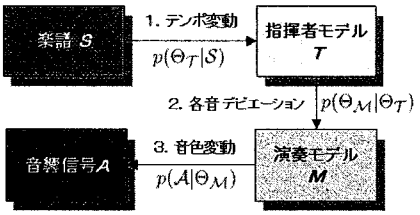

図 3 楽譜からの音楽音響信号生成モデル: 楽譜Sに三つの変動が加わり音響信号Aが生成される。

に関する情報は音響信号に含まれ、楽譜には含まれない。これに注目して、我々は「楽譜に上記三種の変動が順次加わることで、音楽音響信号が生成される」という仮説に基づき、以下のモデルを検討する。

この仮説は、実際の演奏の場面で起こる現象に対応付けて説明することができる。合奏の場面を思い浮かべて欲しい。楽譜 (Score; S) を見て、指揮者は演奏全体のテンポをコントロールする。このとき、指揮者のテンポ (Tempo; T) は厳密に楽譜通りではなく、変動を伴う (テンポ変動)。演奏者は指揮者のテンポに合わせて演奏する。しかしその演奏 (Model; $\mathcal{M}$) には意識的あるいは無意識の揺らぎが生じる (各音のデビエーション)。そして楽器から音が発せられ、音響信号 (Audio; A) が生成される。(音量・音色変動)。指揮者がいない場合は、演奏者自身が頭に思い描く演奏のテンポを、指揮者のテンポに対応付け、同様の説明ができる。

以上の解釈で意味付けられた各種変動は、演奏者、指揮者、楽曲が同じであったとしても、厳密には毎回異なることが容易に想像できる。このことから、これらの変動は確率的に発生する変動として捉えられる。また、各種変動は異なる原因によって発生するため、各種変動は独立に発生するものとして扱える。以上より、図 3 に示す「楽譜からの音楽音響信号生成モデル」を考えることができる。このとき指揮者モデルTは、楽譜にテンポ変動が加わったものなのでTのモデルパラメータΘ_Tが推定すべきテンポに対応する。同様に、演奏音モデル$\mathcal{M}$のパラメータ$\Theta_{\mathcal{M}}$が発音時刻・音長・音高のパラメータである。楽譜Sから音響信号Aが生成する確率は、各種変動の独立性の仮定より

$$p(A|S) = p(A|\Theta_{\mathcal{M}})p(\Theta_{\mathcal{M}}|\Theta_T)p(\Theta_T|S) \qquad (1)$$

と表される。

この生成確率を具体的に与えることで、詳細なアラインメントは、確率的逆問題として定式化が可能である。次項からは、各種変動をモデル化し、「楽譜からの音楽音響信号生成モデル」を具体的に構築する。具体的な変動確率をどのようにモデル化するかが、手法の成否を分けることになるが、今回は特にシンプルな問題として、

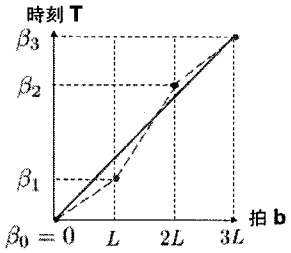

図 4 区分的に一定なテンポモデル: 楽譜通りの場合と (実線) と変動を含む場合 (点線)

- 音高の揺らぎは十分無視できる
- 音長には十分大きい変動を許す

が仮定できる場合を考える。これは、テンポ変動と密接に関係するであろう、発音時刻のデビエーションをモデルとして導入した効果を確認するためである。

3.2 楽譜から指揮者へ – テンポ変動のモデル化

指揮者 (あるいは演奏者の頭の中のテンポ設計) は、本来は音楽的な解釈や意図に基づいて、楽譜通りの一定テンポに対して変動を与える。その変動は、将来は高度な音楽的な楽譜の自動解析と理解から予測できるようになるかも知れないが、当面はそのような知識を仮定せず、従って、確率的な変動と捉えるのが妥当であろう。局所的なテンポ揺らぎを演奏者の各音のデビエーションとして扱えば、指揮者テンポは楽曲のほとんどの場所で区分的に一定であると近似できる。そこで、楽譜上の拍位置bと音響信号上時刻$T(b)$の対応関係を、図 4 に示すような区分線形モデル (固定・等間隔) とする。区間数J、区間長L、テンポが変化する時刻を$\beta_j (j = 1, 2, \ldots, J-1)$とすると、$j$番めの区間の拍と時刻の関係は、

$$T^{(j)}(b) = \frac{\beta_{j+1} - \beta_j}{L} b + ((j+1)\beta_j - j\beta_{j+1}) \qquad (2)$$

で与えられ、これが全てのテンポ情報を表すので、指揮者テンポのモデルパラメータは、$\Theta_T := \{\beta_1, \beta_2, \ldots, \beta_{J-1}\}$である。曲の始め$\beta_0$と、終わり$\beta_J$は固定する。

指揮者によって与えられるテンポは、個人性や楽曲に大きく依存するが、本研究では、汎用のアラインメント手法の構築を念頭に置いて、多くの演奏に共通するテンポの特徴をモデル化する方針をとる。今回は、「テンポ変動は滑らかである」という仮定に基づき、「隣り合う区間のテンポの差は平均が 0 で、分散の小さい正規分布に従う」とする。これによって、テンポ変動の確率は以下で与えられる。

$$p(\Theta_T|S) = \prod_{j=0}^{J-1} \frac{1}{\sqrt{2\pi}\sigma_t} e^{-\frac{(\beta_j - \beta_{j+1}/2 - \beta_{j-1}/2)^2}{2\sigma_t^2 L^2}} \qquad (3)$$

3.3 指揮者から演奏へ – デビエーションのモデル化

指揮者のテンポからの発音時刻の揺らぎは、演奏によっ

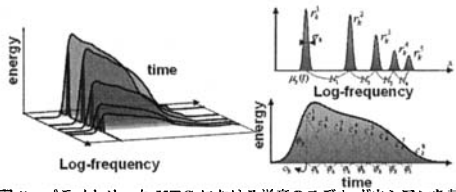

図 5 パラメトリック HTC における単音のモデル:ガウシアンを基底関数として時間・周波数方向に連ねる

て生じる誤差として捉えることができる。「発音時刻 τ_k は、テンポに従う理想時刻 $T(b_k)$ を中心とする分散の小さな正規分布に従う」ことを仮説として、

$$p(\Theta_{\mathcal{M}}|\Theta_T) = \prod_{j=0}^{J-1}\prod_{k_j=1}^{K_j} \frac{1}{\sqrt{2\pi\sigma_{on}^2}} e^{-\frac{(\tau_{k_j}-T^{(j)}(b_{k_j}))^2}{2\sigma_{on}^2}} \tag{4}$$

を各音のデビエーションの確率とする。j は区間のインデックス, k_j, K_j は、区間 j に含まれる、単音のインデックスと、個数である。

3.4 演奏音から音響信号へ – 音色変動のモデル化

演奏音モデル ($\mathcal{M}$) からの音響信号 ($\mathcal{A}$) 生成モデルに基づく優れた音響信号解析手法として、Kameoka らの提案する時間調波構造クラスタリング (Harmonic Temporal Clustering; HTC)[6] がある。

HTC は、多重音解析を、時間周波数平面上における演奏音モデルの配置問題として捉える点が特徴である。この捉え方は、時間周波数各方向のデビエーションを統一的・直感的に扱うのに適しており、詳細なアラインメントにも有効なアイディアである。本稿では、HTC と同様に詳細なアラインメントを時間周波数平面上で問題を扱うために、その第一歩として HTC で用いられているの演奏音モデルと音色変動の確率を用いる。

Kameoka らは、時間-対数周波数上で、単音のパワースペクトルモデルを与え、その和によって演奏音全体のモデルを表した。特に、単音のモデルとして、自由度の高い分布形状を比較的少量のパラメータで表現可能なモデルとして、下式 q_k が提案されている (図 5)。x, t は各々対数周波数と時刻である。

$$q_k(x,t) = \sum_{n,y} S_{k,n,y}(x,t), \tag{5}$$

$$S_{k,n,y}(x,t) \triangleq \frac{w_k v_{k,n} u_{k,n,y}}{2\pi\sigma_k\phi_{k,n}} \times \exp\left(-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_{k,n})^2}{2\phi_{k,n}^2}\right) \tag{6}$$

各パラメータの物理的意味は表 1 に示す通りである。なお、$v_{k,n}, u_{k,n,y}$ のスケールの任意性を取り除くため以下の制約が設けられている。

$$\forall k, \sum_n v_{k,n} = 1, \quad \forall k, \forall n; \sum_y u_{k,n,y} = 1. \tag{7}$$

表 1 単音モデル 各パラメータの物理的意味

表記	物理的意味
$\mu_k(t)$	ピッチ周波数
w_k	合計エネルギー
$v_{k,n}$	n 倍音のエネルギー比
$u_{k,n,y}$	n 倍音における時間方向エネルギー比
τ_k	発音時刻
$Y_k\phi_k$	音長 (Y_k: 定数)
σ_k	周波数方向のひろがり

演奏音のモデルが与えられたので、次は、音色変動の確率を具体的に与える。観測スペクトログラム $W(x,t)$ は、以下の区間

$$D = \left\{x, t \in \mathbb{R} \mid X_0 \leqq x \leqq X_1, T_0 \leqq t \leqq T_1\right\} \tag{8}$$

で定義されているとする。これに対して、演奏音のモデルは、単音モデルの和

$$Q(x,t) = \sum_k q_k(x,t) \tag{9}$$

で表される。このとき、「観測信号のパワースペクトルは、パワースペクトルモデル Q から、時間周波数毎に独立な連続ポアソン分布 (パワースペクトルの非負制約を満たす) に従った揺らぎを持って生ずる」と仮定すると

$$p(\mathcal{A}|\Theta_{\mathcal{M}}) \triangleq C \times \exp\left(\iint_D W(x,t)\log Q(x,t) - Q(x,t)dxdt\right) \tag{10}$$

が得られる[7]。C はパラメータ $\Theta_{\mathcal{M}}$ によらない正規化定数である。

以上のように、三つの変動をモデル化することで「楽譜からの音響信号生成モデル」が具体的に与えられた。

4. 詳細なアラインメントの定式化とその解法

4.1 詳細なアラインメント問題の定式化

前節で与えられた楽譜からの音響信号生成モデルに基づき、本稿で扱う詳細なアラインメント問題は、「楽譜$\mathcal{S}$, 音響信号$\mathcal{A}$を与えられたとき、テンポと演奏のパラメータ $\Theta \triangleq (\Theta_{\mathcal{M}}, \Theta_T)$ を推定する」事後確率最大化の問題として与えられる。ベイズの定理と各変動の独立性より、

$$\begin{aligned}\Theta &= \operatorname*{argmax}_{\Theta} p(\Theta|\mathcal{S},\mathcal{A}) \\ &= \operatorname*{argmax}_{\Theta}\Big(\log p(\mathcal{A}|\Theta_{\mathcal{M}}) \\ &\qquad + \log p(\Theta_{\mathcal{M}}|\Theta_T) + \log p(\Theta_T|\mathcal{S})\Big) \\ &= \operatorname*{argmax}_{\Theta}\Big(J_{A|M} + J_{M|T} + J_{T|S}\Big)\end{aligned} \tag{11}$$

である。ただし、$J_{A|M}, J_{M|T}, J_{T|S}$ は、前節で設計した各種変動の確率分布の対数を取り、パラメータに依存する項だけを取り出したものである。

$$J_{A|M} = \iint_D \Big(W(x,t)\log Q(x,t) - Q(x,t)\Big)dxdt \tag{12}$$

$$J_{M|T} = -\sum_j \sum_{k_j} \frac{(\tau_{k_j} - T(b_{k_j}))^2}{2\sigma_{on}^2} \tag{13}$$

$$J_{T|S} = -\sum_j \frac{(\beta_j - \beta_{j+1}/2 - \beta_{j-1}/2)^2}{2\sigma_t^2} \tag{14}$$

4.2 演奏とテンポの反復推定アルゴリズム

上述の問題に対し、最適な$\Theta_{\mathcal{M}}$, Θ_Tを一挙に解析的に求めることはできないが、一方を固定すれば、解析的に、あるいは安定性のある反復法で、もう一方に関して目的関数を単調増加させられる。これを利用して、以下に示すアルゴリズムにより詳細なアラインメントが行える。

1. $\Theta_{\mathcal{M}}$, Θ_Tに適当な初期値を与える
2. Θ_Tを固定、$J_{A|M} + J_{M|T}$ を$\Theta_{\mathcal{M}}$について増加させる
   (1) 分配関数を式 (17) に従って更新
   (2) $\Theta_{\mathcal{M}}$を式 (18)-(24) に従い更新
3. $\Theta_{\mathcal{M}}$を固定、$J_{M|T} + J_{T|S}$ をΘ_Tについて最大化する
4. 目的関数が収束したら終了, それまで 2,3 を繰り返す

アルゴリズムの各ステップで目的関数の値は単調非減少、目的関数は上に有界なので、局所最適解への収束性が保証される。以下で、Step 2, 3 におけるパラメータの更新方法を示す。

4.3 Step 2: $\Theta_{\mathcal{M}}$の更新

分配関数と呼ばれる新たな変数 $m_{k,n,y}(x,t)$ を導入する。これは、観測スペクトルの一部

$$\ell_{k,n,y}(x,t) \triangleq m_{k,n,y}(x,t) W(x,t) \tag{15}$$

が、部分単音モデル $S_{k,n,y}$ に由来するという解釈を与えるものであり、

$$\sum_{k,n,y} m_{k,y,y}(x,t) = 1,$$

$$\forall k,n,y: \ 0 < m_{k,n,y}(x,t) < 1$$

を満たす。これを用いて、Jensen の不等式を用いると、

$$J_{A|M} \geq \sum_{k,n,y} \iint_D \left(\ell_{k,n,y}(x,t) \log \frac{S_{k,n,y}(x,t)}{m_{k,n,y}(x,t)} - Q(x,t) \right) dxdt \triangleq \tilde{J}_{A|M}$$

$$\left(\text{等号成立は、} \hat{m}_{k,n,y}(x,t) = \frac{S_{k,n,y}(x,t)}{\sum_{k,n,y} S_k(x,t)} \right) \tag{16}$$

が得られる。これによって、

$$\operatorname*{argmax}_{\Theta_{\mathcal{M}}} \left(J_{A|M} + J_{M|T} \right) = \operatorname*{argmax}_{\Theta_{\mathcal{M}}} \left(\max_{m_{k,n,y}(x,t)} \tilde{J}_{A|M} + \log p(\Theta_{\mathcal{M}}|\Theta_T) \right)$$

であるから、本ステップの計算は、式 (16) により、$\tilde{J}_{A|M}$ を $m_{k,n,y}(x,t)$ について最大化した後、さらに全体を$\Theta_{\mathcal{M}}$について最大化することで可能である。後段は、制約式 (7) を考慮して、

$$\tilde{J}_{A|M} + J_{M|T} - \sum_k \left(\gamma_v^{(k)} \sum_n v_{k,n} - 1 \right) - \sum_k \left(\sum_n \gamma_u^{(k,n)} \left(\sum_y u_{k,n,y} - 1 \right) \right) \tag{17}$$

を各パラメータに関して、偏微分し極大値を達成する値に更新することで行われる。今回考えるアラインメント問題で必要なパラメータは発音時刻と音長のみであるが、同時に他のパラメータも更新することで、各音の音色情報も同時に推定することができる。反復回数の添え字を i として、全パラメータの更新式を以下に示す。

$$w_k^{(i)} = \sum_{n,y} \iint_D \ell^{(i)}(x,t) dxdt \tag{18}$$

$$\mu_{k,0}^{(i)} = \frac{1}{w_k^{(i)}} \sum_n \iint_D \ell^{(i)}(x,t)(x - \log n) dxdt \tag{19}$$

$$v_{k,n}^{(i)} = \frac{1}{w_k^{(i)}} \sum_y \iint_D \ell^{(i)}(x,t) dxdt \tag{20}$$

$$u_{k,y}^{(i)} = \frac{1}{w_k^{(i)}} \sum_n \iint_D \ell^{(i)}(x,t) dxdt \tag{21}$$

$$\tau_k^{(i)} = \frac{\frac{T(b_k)}{\sigma_{on}^2} + \frac{1}{\phi_k^2} \sum_{n,y} \iint_D \ell^{(i)}(x,t)(t - y\phi_k^{(i-1)}) dxdt}{1/\sigma_{on}^2 + w_k^{(i)}/\phi_k^2} \tag{22}$$

$$\phi_k^{(i)} = \frac{\sqrt{{a_k}^2 + 4(w_k^{(i)} + 1)b_k} - a_k}{2w_k^{(i)}} \tag{23}$$

$$\begin{cases} a_k^{(i)} &= \sum_{n,y} \iint_D \ell^{(i)}(x,t) y(t - \tau_k) dxdt \\ b_k^{(i)} &= \sum_{n,y} \iint_D \ell^{(i)}(x,t)(t - \tau_k)^2 dxdt \end{cases}$$

$$\sigma_k^{(i)} = \sqrt{\frac{\sum_{n,y} \iint_D (x - \mu_{k,0}^{(i)} - \log n)^2 \ell^{(i)}(x,t) dxdt}{w_k^{(i)}}} \tag{24}$$

4.4 Step 3: Θ_Tの更新

このステップにおける目的関数 $J_{M|T} + J_{T|S}$ をテンポのパラメータで偏微分したものが 0 である、という方程式は、連立一次方程式である。これを解くことで、最大化が行える。

5. 実　　験

本節では、今回提案する手法を実際の楽曲に適用した例を示し、本手法の定性的振舞を観察する。

5.1 実験条件

入力音響信号には、RWC 研究用音楽データベースより RWC-MDB-C-2001(クラシック) より、No.39(C. フランクのヴァイオリンソナタ)・No.29(R. シューマンのトロイメライ: ピアノ独奏曲) から抜粋して利用した。楽譜情報には、同じく RWC 研究用音楽データベースの対応する MIDI ファイル (ほぼアラインメントが取られている) を楽譜通りの発音拍と音価に修正して利用した。

演奏音のモデルは、楽譜の通りの配置から音響信号の長さに合わせた線形伸縮をして初期配置とし、テンポの初期値は、平均のテンポ (α とする) とした。また、分析条件は、テンポ一定の区間を 1 秒、発音時刻・テンポの標準偏差をそれぞれ $\sigma_{on} = 0.25$ 秒, $\sigma_t = \alpha/20$ とした。

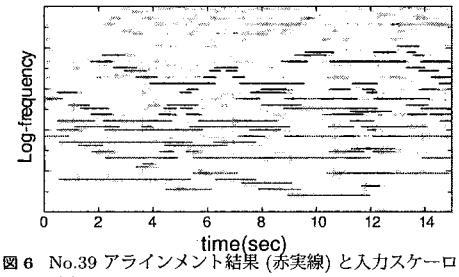

図 6 No.39 アラインメント結果 (赤実線) と入力スケーログラム

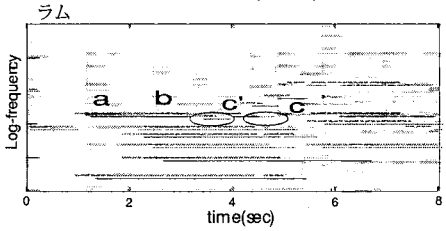

図 7 No.29 アラインメント結果 (実線) と楽譜 (太い点線) と入力スケーログラム

5.2 結　　果

発音時刻の定義については十分な議論が必要だが、今回は定量的評価は行わないため、暫定的に、各演奏音モデルの $\tau_k - \phi_k$ を発音時刻とした (立上り部分の振幅が、最大振幅の約 1/4 に達した地点」に相当)。

No.39 は、テンポがほぼ一定の演奏であるが、アラインメント結果においても、推定テンポは一定 (図 8 左) で、各音の発音時刻・音長の揺らぎは、確かにデビエーションとして推定が行われている (図 6)。

No.29 は、テンポ変動が豊富な演奏である。図 7 に示す通り、楽譜 (点線) と比べて、a-b の発音時刻間隔はテンポ通りだが、音 b だけが遅く演奏される。今回のモデルは発音時刻からテンポを推定するため、この部分のテンポ推定が正しく行えず、その後音 c 以降のアラインメントが (c' が推定結果) 真値と一致しなかった。同様に、区間に置ける発音の数が少ないためテンポ推定が誤っている現象は 7 秒前後でより顕著な形で起きている (図 4 右参照)。この音 c の発音時刻の推定値は、一回の反復毎にゆっくりと真値に近付いてはいたものの、その変化は極めて小さく、仮にさらなる反復によって真値へ収束するとしても実用的ではない。発音時刻変動のパラメータを反復回数や区間によって変える、音長を考慮にいれたテンポ推定を行う等、デビエーションの扱いについて検討の余地が有るだろう。

6. おわりに

本稿では、楽譜情報を利用した音楽音響信号の加工等に必要な、楽譜と音響信号を一音毎対応付ける詳細なアラインメントについて議論した。「楽譜からの音楽音響信号生成モデル」に基づいた MAP 推定によるアラインメントを提案し、適用例から、確かに微小変動を扱いうる枠組であるが、音長のモデル化等まだ検討の余地が有ることを確認した。

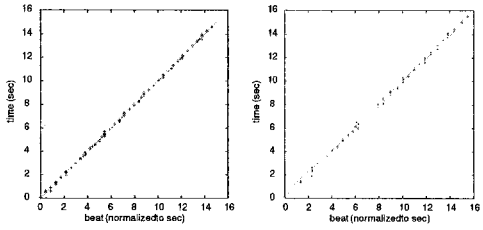

図 8 推定テンポ (緑点線の区分直線) と、各音発音時刻 (赤点): (左)No.39 (右)No.29

本稿で示したモデル化の例は選択肢の一つに過ぎない。今回は、各音のデビエーションにシンプルなモデルを与えたが、音長・音高の揺らぎもモデル化が可能である。今後は、より適切なモデル構築、手法の定量的な評価実験とともに、大域的最適解への収束を保証するアルゴリズムを検討を行う。今回の枠組は、アラインメントをとると同時に各音の音色情報の推定が可能であった。その情報を利用し除去にも取り組みたい。

謝　　辞

本研究の一部は、科学技術振興機構 CREST 研究課題「時系列メディアのデザイン転写技術の開発」として行われた。また、亀岡弘和氏 (NTT CS 研) には、有益なコメントを頂いた。

参 考 文 献

1) K.Itoyama et al., "Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals," *Proc. of IEEE ICASSP*, Hawaii, April, 2007.
2) N.Orio et al., "Alignment of Monophonic and Polyphonic Music to a Score," *Proc. of ICMC2001*, pp.155–158, 2001.
3) N.Orio et al., "Score Following Using Spectral Analysis and Hidden Markov Models," *Proc. of ICMC2001*, pp.151–154, 2001.
4) C.Raphael, "A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores," *Proc. of the 5th ISMIR*, pp.387–394, 2004.
5) 松本他, "合奏音楽音響信号からの 1 パート除去の検討, " 日本音響学会春季研究発表会講演集, 3-7-8, pp.749–750, Mar, 2007.
6) H.Kameoka et al., "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 15, No. 3, pp.982–994, Mar, 2007.
7) J.Le Roux et al., "Single and Multiple F0 Contour Estimation Through Parametric Spectrogram Modeling of Speech in Noisy Environments," *IEEE Trans. on Audio, Speech and Language Processing*, Vol.15, No.4, pp.1135–1145, May, 2007.