Vol.2011-MUS-91 No.19
2011/7/29

MAHL:演奏者間のインタラクション分析のためのスコアアライメント手法の提案

前澤 陽[†1,*1] 糸山克寿
尾形哲也 奥乃博

本稿では、楽器パート毎に、楽譜と音響信号のアライメントを算出する手法を提案する。本手法では、各楽器パートに共通の、自己回帰過程に従うテンポモデルを持たせる。各楽器パートの時系列は隠れセミマルコフモデルに従い、状態継続長の事前分布としてテンポモデルを持つ。また、音響信号の出力は潜在的調波配分法に従う。パート間の揺らぎを持たせない場合の、アライメントの性能を評価し、アライメント手法としての有用性が確認された。また、演奏における発音タイミングの揺らぎがモデル化できることが示唆された。

MAHL:Score Alignment Method for Analyzing Inter-performer interaction

Akira Maezawa,[†1,*1] Katsutoshi Itoyama,
Tetsuya Ogata and Hiroshi G. Okuno

This paper presents a method to align an audio signal and individual music instrument parts comprising a music score. Such method allows a machine to analyze temporal interaction of music performers. Proposed method is based on fitting multiple Hidden Semi-Markov Models (HSMM) to the observed audio signal, each HSMM of which emits Latent Harmonic Allocation parameters. Each HSMM corresponds to a music instrument part, and the state duration probability is conditioned on an auto-regressive tempo model. Evaluation suggests usefulness as score alignment method, and hints at the usefulness as multiple part alignment method.

1. はじめに

膨大なクラシック音楽のデータ集から、自分好みの演奏を見つけるのは難しい。クラシック音楽は、単一の楽譜を、様々な演奏者がそれぞれの解釈に基づき演奏する。つまり、楽譜には、演奏としての音響信号を定義するのに十分な情報は含まれておらず、音響信号には演奏者の解釈が常に介在する。そのため、クラシック音楽の視聴において、演奏者の解釈の違いを楽しむのは非常に重要な要素である。一方で、100 年以上の歴史を持つ膨大なクラシック音楽の録音のデータから、リスナー好みの解釈を見つけるのは困難である。そこで、計算機を用い、音楽の解釈を推定することにより、クラシック音楽検索に応用する必要性が考えられる。そこで、我々は、音楽の解釈の違いは音色、音量バランス、テンポ、繰り返し記号などの選択、といったものにあるという観点から、これらを統合した、音楽解釈のモデルを構築することを目指す。

テンポの取り扱い方は、従来、楽譜表現と音響信号の時間的対応付けを取る、スコアアライメントという手法により実現された。しかし、スコアアライメント手法では、演奏者間のタイミングのばらつきを推定することができない。というのも、従来のスコアアライメント手法[1)–4)] は、楽譜内において同時刻に発音される音は、音響信号でも同様に同時刻に発音されるという前提があるからだ。演奏者間のばらつきは、テンポの変動が激しい曲や小規模のアンサンブルでは、音楽表現において重要な役割を果たしていると考えられる。よって、このような楽曲において、スコアアライメントが果たす役割は不十分であり、演奏されているパート毎のアライメントが必要であると考えられる。

そこで、我々は、個別の楽器に対するアライメントを算出する手法、MAHL[⋆1]を提案する。MAHL では楽曲の時系列を、演奏者間が共通のものとして同意するテンポと、実際に各パートが奏でる時系列の二種類が混在するものとしてモデル化する。前者は、楽器パート共通のテンポモデル、そして後者は楽器パート分だけ存在する隠れセミマルコフモデル (HSMM) が用いられる。テンポは前後の連続性が保たれるよう自己回帰過程に従う。また、HSMM はそれぞれ、調波音のベイズ的モデルである LHA[5)] のパラメータを出力する。

†1 京都大学情報学研究科
Graduate School of Informatics, Kyoto University
∗1 現在，ヤマハ株式会社
Presently with Yamaha Corporation
⋆1 **M**ultiple **A**utoregressive-tempo **H**SMM with **L**HA emission

2. MAHL モデルの定式化

本手法は、入力信号の定 Q 変換に対し、入力された楽譜表現とのアライメントを行う。以後、楽譜において、特定の楽器 h が奏でている特有の音高 i の対を**楽器音高ペア** (h,i) と呼ぶ。すなわち、楽譜の特定の位置は複数の楽器音高ペアの集合であり、楽譜とはこれらを連結したものである。

まず、本手法は、観測されたパワーが、単一の楽器から生成されたものとする。また、単一の時間周波数ビンに含まれるすべてのパワーは単一の楽器が生成したものとする:

$$p(Z^{(m)}|m) = \prod_{t,f,h} m_h(f,t)^{Z_h^{(m)}(f,t)X(f,t)} \tag{1}$$

$$p(m|m_0) = \prod_{t,f} \mathrm{Dir}\left(m(f,t)|m_0(f,t)\right) \tag{2}$$

ここで、$m_h(f,t)$ は楽器 h が時間周波数ビン (f,t) を生成する尤度であり、演奏者の時間周波数マスクと考えることができる。$m_0(f,t)$ はその事前分布のハイパーパラメータであり、今回は無情報 $(=1)$ に設定する。

次に、音色と音量の選定にアライメントの性能が左右されない、パート毎の CQT の生成モデルを考える。音量と音色のロバストネスを実現するために、スペクトルを潜在的調波配分法 (LHA) を用いてモデル化する。LHA の出力は、現在の楽譜位置に依存する。各時間フレームにおけるスペクトルは LHA に従い生成されると仮定する。 ただし、LHA の定式化と違い、調波構造は楽器音高ペア内で共有されているとし、また音量バランスは音符内で一貫していると仮定する。さらに、ある楽器の状態内に置ける周波数ビンは単一の楽器の、単一の倍音から生成されるとする。$Z_i^{(i)}(h,f,d)$ を、状態 d において楽器音高ペア (h,i) が周波数 f が占拠している場合 1 でそれ以外は 0 の二値行列とし、$Z_j^{(h)}(f,h,i)$ を、周波数 f が、楽器音高ペア (h,i) の第 j 倍音から生成される場合 1 の二値行列とする。$Z_{l,d}^{(s)}(h,t)$ は時刻 t が、楽器 h が状態 d にいて、次の状態に遷移するまでのフレーム数が l のとき 1 の値をとる二値行列とする。楽器音高ペア (h,i) の基本周波数が $\mu_{h,i}$ であり、 窓関数の影響などにより分散 $\lambda_{h,i}^{-1/2}$ で隣接する周波数でパワーが観測されるとする。以上より、観測信号の尤度は次のように表すことができる:

$$p(X|Z^{(i,h,s)},\mu,\lambda) = \prod_{t,h,i,j,f,d,l} \mathcal{N}\left(\log f/j|\mu_{h,i},{\lambda_{h,i}}^{-1}\right)^{Z_{l,d}^{(s)}(h,t)X(f,t)Z_i^{(i)}(h,f,d)Z_j^{(h)}(f,h,i)Z_h^{(m)}(f,t)} \tag{3}$$

調波構造と音量バランスは多項分布に従うと仮定する。

$$p(Z^{(i)}|E,Z^{(s)}) = \prod_{t,h,i,f,d,l} e_i(h,d)^{Z_{l,d}^{(s)}(h,t)X(f,t)Z_i^{(i)}(h,f,d)Z_h^{(m)}(f,t)} \tag{4}$$

$$p(Z^{(h)}|A,Z^{(i,s)}) = \prod_{t,h,i,j,f,d,l} a_j(h,i)^{Z_{l,d}^{(s)}(h,t)X(f,t)Z_i^{(i)}(h,f,d)Z_j^{(h)}(f,h,i)Z_h^{(m)}(f,t)} \tag{5}$$

e と a をそれぞれ**音符生起確率**と**倍音生起確率**と呼ぶ。これらは、音符の相対音量と倍音ピークの相対強度にそれぞれ対応すると考えることができる。これらを更に確率変数としてして扱い、事前分布を無情報にすることで、音色と音量の変化に対するロバストネスを実現できると考えられる。そこで、音符生起確率と倍音生起確率の事前分布としてディリクレ分布をおき、基本周波数の事前分布として Normal-Gamma 分布を置く:

$$p(\mu,\lambda|\nu,b,m,l) = \prod\nolimits_{h,i} \mathcal{NG}(\mu_{h,i},\lambda_{h,i}|m_{h,i}^{(H)},b_{h,i}^{(H)},l_{h,i}^{(H)},\nu_{h,i}^{(H)}) \tag{6}$$

$$p(E|E_0) = \prod\nolimits_{h,d}^{H,D} \mathrm{Dir}(e(h,d)|e_0(h,d)) \tag{7}$$

$$p(A|A_0) = \prod\nolimits_{h,i}^{H,I} \mathrm{Dir}(a(h,i)|a_0(h,i)) \tag{8}$$

楽譜時系列 $Z^{(s)}$ の分布として HSMM を仮定する。 初期状態の確率分布を π とする。

$$p(Z^{(s)}|T,\pi,\tau) = \prod_h \pi^{Z^{(s)}(h,1)} \prod_{t=2,l,d,d'\neq d} \left(\tau_{d'}(d)\mathcal{N}\left(\log\frac{l}{\mathcal{L}_d}|T_d,\sigma_T^2\right)\right)^{Z_{1,d'}^{(s)}(h,t-1)Z_{l,d}^{(s)}(h,t)} \tag{9}$$

$$p(\pi|\pi_0) = \mathrm{Dir}(\pi|\pi_0) \tag{10}$$

$$p(\tau|\tau_0) = \prod_d \mathrm{Dir}(\tau(d)|\tau_0(d)) \tag{11}$$

式 (9) は、楽譜時系列を、拍長と楽譜上の状態遷移の組み合わせとして表すことを意味する。τ は、複雑な楽譜構造を HMM のように記述できる。T_d は楽譜位置 d における対数拍長である。T_d の連続性を保たせると、音楽的に妥当な拍長のモデル化が可能となる。そこで、T_d を平滑化させるために、LDS をおく:

$$p(T) = \prod_d \mathcal{N}\left(T_d|T_{d-1},\mathcal{L}_{d-1}{\lambda^{(T)}}_d^{-1}\right) \tag{12}$$

$$p(\lambda^{(T)}) = \prod_d \mathcal{G}({\lambda^{(T)}}_d|l_d^{(T)},\nu_d^{(T)}) \tag{13}$$

本手法では、これらの事後分布を推定し、状態系列 $Z^{(s)}$ を音価 l に対して積分消去したものの事後確率を最大化させる状態系列 $\arg\max\sum_l Z_{l,d}^{(s)}(t)$ をスコアアライメントとする。しかし、事後分布の推定は困難であるため、変分近似に基づく EM アルゴリズム (VBEM) を用いて事後分布を推定する。VBEM では、事後分布 $q(MASK,LDS,LHA,HSMM)$ が $q_{MASK}(MASK)q_{LDS}(LDS)q_{LHA}(LHA)q_{HSMM}(HSMM)$ と因子分解できると仮定する。このような分布を変分事後分布と呼ぶ。ここで、$q_{HSMM}(HSMM) = \prod_h q_{Z^{(s)}(h)}(Z^{(s)}(h))q_\pi(\pi)q_\tau(\tau)$ と因子分解でき、$q_{LHA}(LHA) = q_{Z^{(h)}}(Z^{(h)})q_{Z^{(i)}}(Z^{(i)})q_{\mu,\lambda}(\mu,\lambda)$ と因子分解でき、$q_{MASK}(MASK) = q_{Z^{(m)}}(Z^{(m)})q_m(m)$ と因子分解でき、$q_{LDS}(LDS) = q_T(T)q_{\lambda^{(T)}}(\lambda^{(T)})$ と因子分解できるとする。変分事後分布の推定は、同時分布との KL ダイバージェンスの最小化問題として定式化できる。すると、任意の因子 Z は、以下のように更新できる。

$$q_Z(Z) \propto \exp\left\langle \log p(X,LDS,LHA,MASK,HSMM)\right\rangle_{\neg Z} \tag{14}$$

ただし、$\langle f(x)\rangle_x$ は x の下での $f(x)$ の期待値であり、$\neg y$ とは、y 以外のすべての確率変数のことを指す。推定は、KL ダイバージェンスが収束するまで、各確率変数の変分事後分布を交互に更新する。**図 1** にグラフィカルモデルを図示する。

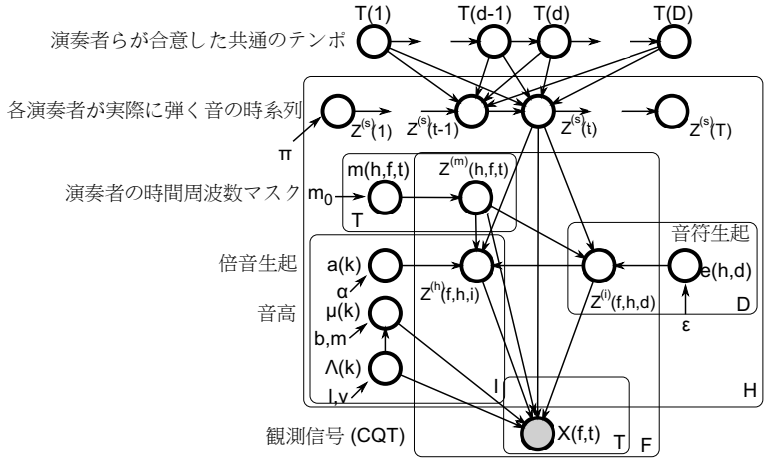

図 1 MAHL のグラフィカルモデル.

3. モデルの推論

簡単のため、次の変数を定義する:

$$l\mathcal{N}_f(h,i,j) = \left\langle \log \mathcal{N}\left(\log f/j | \mu_{h,i}, {\lambda_{h,i}}^{-1}\right)\right\rangle \tag{15}$$
$$= -\frac{1}{2}\left(\frac{\bar{l}_{h,i}}{\bar{\nu}_{h,i}}(\log f/j - \bar{m}_{h,i})^2 + \frac{1}{\bar{b}_{h,i}}\right) - \log 2\pi\bar{\nu}_{h,i} + \psi(\bar{l}_{h,i})$$
$$el(d,l) = \langle \log p(\log l | T_d)\rangle_T \tag{16}$$
$$\eta_d(h,t) = \sum_l \left\langle Z^{(s)}_{d,l}(h,t)\right\rangle_{Z^{(s)}} \tag{17}$$
$$lA_j(h,i) = \langle \log a_j(h,i)\rangle_{a(h,i)} = \psi(\bar{\alpha}_j(h,i)) - \psi\left(\sum_{l=1}^{M}\bar{\alpha}_l(h,i)\right) \tag{18}$$
$$lE_i(h,d) = \langle \log e_i(h,d)\rangle_{e(h,d)} = \psi(\bar{\epsilon}_i(h,d)) - \psi\left(\sum_{l=1}^{K}\bar{\epsilon}_l(h,d)\right) \tag{19}$$
$$lM_h(f,t) = \langle \log m_h(f,t)\rangle_{m(f,t)} = \psi(\bar{m}_{0h}(f,t)) - \psi\left(\sum_{l=1}^{H}\bar{m}_{0l}(f,t)\right) \tag{20}$$
$$e\tau'_d(d) = \langle \log \tau_{d'}(d)\rangle_{\tau(d)} = \psi(\bar{\tau}'_d(d)) - \psi\left(\sum_{l=1}^{D}\bar{\tau}_l(d)\right) \tag{21}$$

ここで、$\psi(x)$ はディガンマ関数である。また、$\langle f(x)\rangle_x$ は、確率変数 x の下での関数 $f(x)$ の期待値である。

3.1 時間周波数マスクの変分 E ステップ

観測音を各パートの音に割り当てる時間周波数マスクを以下のように更新する:

$$q_{Z^{(m)}}(Z^{(m)}) = \prod_{h,t,f}{\gamma_h^{(m)}(f,t)}^{Z_h^{(m)}(f,t)} \tag{22}$$

ただし、$\gamma_h^{(m)}(f,t) = \frac{\rho_h^{(m)}(f,t)}{\sum_{h'}\rho_{h'}^{(m)}(f,t)}$ であり、$\rho_h^{(m)}(f,t)$ は次のように与えられる:

$$\log \rho_h^{(m)}(f,t) = lM_h(f,t) + \sum_{d,i} X(f,t)\eta_d(h,t)\left[lE_{i,h}(d) + \sum_j \xi_j(f,h,i)\left(l\mathcal{N}_f(h,i,j) + lA_j(h,i)\right)\right] \tag{23}$$

3.2 LHA の変分 E ステップ

HSMM の各状態 d における、楽器音高ペア (h,i) が周波数 f に占める割合 $Z^{(i)}$ を次のように更新する:

$$q_{Z^{(i)}}(Z^{(i)}) = \prod_{h,i,f,d}{\gamma_i(h,f,d)}^{Z_i^{(i)}(h,f,d)} \tag{24}$$

ただし

$$\gamma_i(h,f,d) = \frac{\rho_i(h,f,d)}{\sum_i \rho_i(h,f,d)} \tag{25}$$

であり、 ρ は次のように表されるとする:

$$\log \rho_i(h,f,d) = \left(\sum_t \gamma_h^{(m)}(f,t)X(f,t)\eta_d(h,t)\right) \times \left[lE_{i,h}(d) + \sum_j \xi_j(f,h,i)\left(l\mathcal{N}_f(h,i,j) + lA_j(h,i)\right)\right] \tag{26}$$

式 (26) 右辺第 1 項を状態 d で重み付けたスペクトルの周波数平均、また第 2 項を音符 i における音量の対数期待値と、音符 i の調波構造と倍音ピークの対数期待値による重み付けと考えると、$\rho_i(f,d)$ は状態 d 内の平均スペクトルを、音符ごとに分配するとみなすことができる。

同じように、各楽器音高ペア i における、倍音 j が周波数 f に占める割合 $Z^{(h)}$ を次のように更新する:

$$q_{Z^{(h)}}(Z^{(h)}) = \prod_{j,i,f,h}{\xi_j(f,h,i)}^{Z_j^{(h)}(f,h,i)} \tag{27}$$

ただし $\xi_j(f,h,i) = \frac{\phi_j(f,h,i)}{\sum_k \phi_k(f,h,i)}$ であり、ϕ は次のように表される:

$$\log \phi_j(f,h,i) = \left(\sum_{t,d} \gamma_h^{(m)}(f,t)X(f,t)\eta_d(h,t)\gamma_{i,h}(f,d)\right) \times \left[l\mathcal{N}_f(h,i,j) + lA_j(h,i)\right] \tag{28}$$

式 (28) も式 (26) と似たように、楽器音高ペア i 内の平均スペクトルを倍音毎に分配するものとみなせる。

3.2.1 LHA における変分 M ステップ

楽器音高ペアの独立性により $q_E = \prod_d q_{e(d)}$ と表せられる。 また、多項分布とディリク

レ分布の共役性により、 $e(d)$ の事後分布は次のように求められる:

$$q_{e(d)} \sim \mathrm{Dir}\left(e(d)|\epsilon(\bar{d})\right) \tag{29}$$

ここで、$\bar{\epsilon}_{i,h}(d) = e_{0\,i,h}(d) + \sum_{t,f} \eta_d(h,t)\,\gamma_h^{(m)}(f,t)X(f,t)\gamma_{i,h}(f,d)$ と与えられる。同じく、倍音生起確率も $q_A = \prod_i q_{a_{i,h}}$ と表すことができ、$a_{i,h}$ の事後分布は次のように求められる:

$$q_{a_{i,h}} \sim \mathrm{Dir}\left(a_{i,h}|\alpha_{\bar{i},h}\right) \tag{30}$$

ただし,

$$\bar{\alpha}_{i,h} = a_0(i,h) + \sum_{t,f,d} \gamma_h^{(m)}(f,t)X(f,t)\gamma_{i,h}(f,d)\eta_d(t)\,\xi(f,h,i) \tag{31}$$

である。基本周波数とその分散の事後分布は、基本周波数の独立性から $q_{\mu,\lambda} = \prod_{i,h} q_{\mu_{i,h},\lambda_{i,h}}$ であり、また正規分布と $\mathcal{NG}$ 分布の共役性により、$q_{\mu_{i,h},\lambda_{i,h}}$ は次のパラメータで与えられる $\mathcal{NG}(\bar{m}_{i,h}^{(H)}, \bar{b}_{i,h}^{(H)}, \bar{l}_{i,h}^{(H)}, \bar{\nu}_{i,h}^{(H)})$ である:

$$\bar{m}_{i,h}^{(H)} := \frac{m_{i,h}b_{i,h} + N_{\gamma,i,h}\left\langle \log(f/j)\right\rangle_{\psi(h,i)}}{b_{i,h} + N_{\gamma,i,h}} \tag{32}$$

$$\bar{b}_{i,h}^{(H)} := b_{i,h} + N_{\gamma,i,h} \tag{33}$$

$$\bar{l}_{i,h}^{(H)} := l_{i,h} + N_{\gamma,i,h} \tag{34}$$

$$\bar{\nu}_{i,h}^{(H)} := \nu_{i,h} + \frac{1}{2}\frac{b_{i,h}N_{\gamma,i,h}}{b_{i,h}+N_{\gamma,i,h}}\Bigg(\left(\left\langle \log(f/j)\right\rangle_{\psi(h,i)} - m_{i,h}\right)^2 + N_{\gamma,i,h}^2\left\langle \log(f/j) - \left\langle \log(f/j)\right\rangle_{\psi(i)}\right\rangle_{\psi(h,i)}^2 \Bigg) \tag{35}$$

$\psi_{f,j}(h,i)$ は次のような多項分布である:

$$\psi_{f,j}(i) = \sum_{d,t} \gamma_{i,h}(f,d)\xi_j(f,h,i)\eta_d(h,t)\,\gamma_h^{(m)}(f,t)X(f,t)/N_{\gamma,i,h} \tag{36}$$

$$N_{\gamma,i,h} = \sum_{d,t,f,j} \gamma_{i,h}(f,d)\xi_j(f,h,i)\eta_d(h,t)\,\gamma_h^{(m)}(f,t)X(f,t) \tag{37}$$

3.2.2 LDS の変分 E ステップ

LDS の更新には、カルマンスムーザーと同様、時系列の事後分布を前向き後ろ向きアルゴリズムを用い求める。

$q_T(T)$ は $\psi_l(d) = \sum_{h,t=2,d'\neq d} \zeta_{d'}(t,l,h,d)$ とすると、次のように、カルマンスムーザーと似た形で与えられる:

$$\log q_T(T) = \sum_d \sum_l \psi_l(d)\left\langle \log \mathcal{N}\left(\log\frac{l}{\mathcal{L}_d}|T_d,\sigma_T^2\right)\right\rangle_{\sigma_T^2} + \left\langle \log \mathcal{N}\left(T_d|T_{d-1},\mathcal{L}_{d-1}{\lambda^{(T)}}_d^{-1}\right)\right\rangle_{\lambda^{(T)}{}_d} \tag{38}$$

通常の LDS では、ベクトルを出力するのに対し、本手法では l に対するヒストグラムを出力する点が異なる。前向きアルゴリズムは次のように表される:

$$\begin{aligned}\alpha_d^{(L)}(T_d) &= p(T_d|\psi(1:d)) \\ &\propto \int \alpha_{d-1}^{(L)}(T_{d-1})\,p(T_d|T_{d-1},\mathcal{L}_{d-1}{\lambda^{(L)}}_d^{-1}) \times \textstyle\prod_l p(\log\frac{l}{\mathcal{L}_d}|T_d,\sigma_T^2)^{\psi_l(d)}dT_{d-1} \\ &= \int \mathcal{N}(T_{d-1}|u_{d-1},s_{d-1})\mathcal{N}\left(T_d|T_{d-1},\mathcal{L}_{d-1}{\lambda^{(L)}}_d^{-1}\right)dT_{d-1}\textstyle\prod_l \mathcal{N}\left(\log l/\mathcal{L}_d|T_d,\sigma_T^2\right)^{\psi_l(d)} \\ &= \mathcal{N}(T_d|u_d,s_d)\end{aligned} \tag{39}$$

非積分項の指数の中において T_{d-1} を積分消去し T_d に対し平方完成すると、u_d, s_d は, $m_d = \left(\frac{1}{s_{d-1}} + \frac{\left\langle {\lambda^{(L)}}_d\right\rangle}{\mathcal{L}_{d-1}}\right)^{-1}$ とすると次のように求まる:

$$s_d^{-1} = \textstyle\sum_l \psi_l(d)\frac{1}{\sigma_T^2} + \frac{\left\langle {\lambda^{(L)}}_d\right\rangle}{\mathcal{L}_{d-1}} - m_d\left(\frac{\left\langle {\lambda^{(L)}}_d\right\rangle}{\mathcal{L}_{d-1}}\right)^2 \tag{40}$$

$$u_d = s_d\left(m_d\frac{\left\langle {\lambda^{(L)}}_d\right\rangle}{\mathcal{L}_{d-1}}\frac{u_{d-1}}{s_{d-1}} + \textstyle\sum_l \frac{\psi_l(d)}{\sigma_T^2}\log\frac{l}{\mathcal{L}_d}\right) \tag{41}$$

同様に後ろ向き変数を次のように求める:

$$\begin{aligned}\beta_d^{(L)}(T_d) &= p(\psi(d+1:T)|T_d) \\ &\propto \int p(\psi(d+2:T)|T_{d+1})p(T_{d+1}|T_d) \times \prod_l p(\log\frac{l}{\mathcal{L}_d},T_{d+1})^{\psi_l(i+1)}dT_{d+1} \\ &= \int \beta_{d+1}(T_{d+1})\prod_l \mathcal{N}\left(\log\frac{l}{\mathcal{L}_{d+1}}|T_{d+1},\sigma_T^2\right)^{\psi_l(d+1)} \times \mathcal{N}\left(T_{d+1}|T_d,\mathcal{L}_d{\lambda^{(L)}}_{d+1}^{-1}\right)dT_{d+1} \\ &= \mathcal{N}(T_d|v_d,q_d)\end{aligned} \tag{42}$$

同じく平方完成を行うと次を得る。ただし $n_d = \left(\frac{1}{q_{d+1}} + \frac{\left\langle {\lambda^{(L)}}_{d+1}\right\rangle}{\mathcal{L}_d} + \sum_l \frac{\psi_l(d+1)}{\sigma_T^2}\right)^{-1}$ とする:

$$q_d^{-1} = \frac{\left\langle {\lambda^{(T)}}_{d+1}\right\rangle}{\mathcal{L}_d} - n_d\left(\frac{\left\langle {\lambda^{(T)}}_{d+1}\right\rangle}{\mathcal{L}_d}\right)^2 \tag{43}$$

$$v_d = n_d q_d\frac{\left\langle {\lambda^{(T)}}_{d+1}\right\rangle}{\mathcal{L}_d}\left(\sum_l \frac{\psi_l(d+1)}{\sigma_T^2}\log\frac{l}{\mathcal{L}_{d+1}} + \frac{v_{d+1}}{q_{d+1}}\right) \tag{44}$$

これらを用いて、拍長の事後分布を次のように得る：

$$q(T_d|l_{1:T}) = \alpha_d^{(L)}(T_d)\,\beta_d^{(L)}(T_d) = \mathcal{N}\left(T_d|\frac{1}{q_d^{-1}+s_d^{-1}}\left(\frac{v_d}{q_d}+\frac{u_d}{s_d}\right),\frac{1}{q_d^{-1}+s_d^{-1}}\right) \tag{45}$$

3.3 HSMM の変分 EM ステップ

状態継続長の期待値を次のように得る:

$$\begin{aligned}el(d,l) &= \left\langle -\frac{1}{2\sigma_T^2}\left(\log l/\mathcal{L}_d - T_d\right)^2 - \log\left(2\pi\sigma_T^2\right)\right\rangle_{T_d} \\ &= -\frac{1}{2\sigma_T^2}\left(\log\frac{l}{\mathcal{L}_d} - \frac{1}{q_d^{-1}+s_d^{-1}}\left(\frac{v_d}{q_d}+\frac{u_d}{s_d}\right)\right)^2 - \frac{1}{2\sigma_T^2\left({q_d}^{-1}+{s_d}^{-1}\right)} - \log\left(2\pi\sigma_T^2\right)\end{aligned} \tag{46}$$

状態遷移確率 τ の期待値は次のように求まる:

$$q_\tau = \prod_d \mathrm{Dir}\left(\tau_{d'}(d)|\bar{\tau}_{d'}(d)\right) \tag{47}$$

ただし、

$$\bar{\tau}_{d'}(d) = \tau_{0\,d'}(d) + \sum_{t,h,l} \zeta_{d'}(t,l,h,d) \tag{48}$$

$q_{Z^{(s)}}(Z^{(s)})(h,:)$ は次のように求まる:

$$\begin{aligned}\log q(Z^{(s)}) = \sum_{l,d} Z^{(s)}_{l,d}(h,1)\left\langle \log \pi \right\rangle_\pi + \sum_{t=2,d'\neq d} Z^{(s)}_{1,d'}(h,t-1)Z^{(s)}_{l,d}(h,t)(e\tau_{d'}(d)+el(d,l)) \\ + \sum_{t=1} Z^{(s)}_{l,d}(h,t)\gamma^{(m)}_h(f,t)X(f,t)\log\kappa_d(h,f)\end{aligned} \tag{49}$$

ただし

$$\log\kappa_d(h,f) \;=\; \gamma_{i,h}(f,d) \;\times\; \Big(\sum_{i,j}\xi_j(f,h,i)\big(lN_f(h,i,j) \;+\; lA_j(h,i)\big) \;+\; lE_{i,h}(d)\Big) \tag{50}$$

$\kappa_d(h,f)$ は状態 d が出力する、正規化されていないスペクトルの期待値と解釈できる。LHA の不確定さが高い状態 d の $\kappa_d(h,f)$ の周波数軸の累計は小さな値をとるため、 状態系列の期待値に、LHA がどれだけ信号を説明できるかの良し悪しが影響する。

また、これは通常の HSMM と同じ形をしているため、期待値の計算において前向き後ろ向きアルゴリズムを使用できる。$\alpha^{(H)}$ を HSMM の前向き変数、$\beta^{(H)}$ を後ろ向き変数とすると、次の漸化式が求まる:

$$\begin{aligned}\alpha^{(H)}_{l,d}(h,t) &= p(Z^{(s)}_{(l,d)}(t)=1|X(1)\cdots X(t)) \\ &\propto \textstyle\sum_{l',d'} \alpha^{(H)}_{l',d'}(t)\exp\left\langle \log\tau_{l',d'}(l,d)\right\rangle_\tau \prod_f \kappa_d(h,f)^{\gamma^{(m)}_h(f,t)X(f,t)} \\ &= \big(\textstyle\prod_f \kappa_d(h,f)^{\gamma^{(m)}_h(f,t)X(f,t)}\big)\times\big(\alpha^{(H)}_{l+1,d}(h,t-1) \\ &\qquad + \textstyle\sum_{d'}\exp\left(e\tau_d(d')+el(d,l)\right)\alpha^{(H)}_{1,d'}(h,t-1)\big)\end{aligned} \tag{51}$$

$$\begin{aligned}\beta^{(H)}_{l,d}(h,t) &= p(X_{t+1}(f)\cdots X_T(f)|Z^{(s)}_{(l,d)}(t)=1) \\ &= \textstyle\sum_{l',d'}\beta^{(H)}_{l',d'}(h,t+1)\,e^{\left\langle\log\tau_{l,d}(l',d')\right\rangle_\tau}\prod_f \kappa_{d'}(h,f)^{\gamma^{(m)}_h(f,t)X(f,t+1)} \\ &= \begin{cases}\left(\prod_f \kappa_d(h,f)^{\gamma^{(m)}_h(f,t)X(f,t+1)}\right)\beta^{(H)}_{l-1,d}(h,t+1) & l>1 \\ \sum_{d'}\left(\prod_f \kappa_{d'}(h,f)^{\gamma^{(m)}_h(f,t)X(f,t+1)}\right)\exp\left(e\tau_{d'}(d)\right) & \\ \times\sum_{l'}\beta^{(H)}_{l',d'}(h,t+1)\exp\left(el(d',l')\right) & l=1\end{cases}\end{aligned} \tag{52}$$

これらを用い、次の期待値を求める:

$$\eta_d(h,t) \propto \textstyle\sum_l \alpha^{(H)}_{l,d}(h,t)\,\beta^{(H)}_{l,d}(h,t) \tag{53}$$

$$\zeta_{d'}(t,l,h,d) \propto \alpha^{(H)}_{1,d'}(h,t-1)e^{e\tau_{d'}(d)+el(d)}\beta^{(H)}_{l,d}(h,t) \tag{54}$$

HSMM における前向き後ろ向きアルゴリズムの計算には、膨大なメモリを必要とされるように見える。しかし、$\alpha^{(H)}_{l,d}(h,t)$ の漸化式に着目すると、式は h を固定すればそれ以外の h に対して独立であり、必要な統計量も同じことが言える。よって、H 個の HSMM の前向き後ろ向きアルゴリズムを個別に実行すればよい。また、個別の HSMM を実行する際に、漸化式の結果を $\sqrt{LTD}$ フレームおきに保存して、その間の結果は必要に応じて再計算する。これにより、計算時間は $O(HLTD)$、メモリは $O(\sqrt{LTD})$ になる。

4. 評価実験

まず、本手法をスコアアライメント手法として用いた場合の有用性を評価することにより、音源のモデルや時系列モデルの妥当性を検証する。具体的には、すべてのパートを一つのパートに割り当て、$H=1$ とする。実験では、(1) 自己回帰過程 (LDS) を用いた拍長モデルの有用性、(2) 音色と音量に不確定性を持たせる MAHL を用いることの有用性、の二点を評価する。(1) を評価するために、タイミングモデルに LDS を用いない手法を用意する。音価に比例するような音長の期待値を持った HSMM を用意した (HL)。固定されたテンポに依存するという意味では、このタイミングモデルは[3] と同等である。(2) を評価するために、調波構造と音量バランスに事前分布を持たせないものを用意する (MAP-MAHL)。スペクトルモデルは[1] と同等になる。調波構造のモデルは[1] で用いられた値を使った。サンプリング周波数 8kHz、 分析フレームレート 20 E_0 と A_0 は無情報に設定し、調波構造の事前分布は楽譜に記載された音高を平均とし標準偏差を 20 cent とした。CQT は 0.25 半音毎に評価した。

RWC クラシック音楽データベース[6] 40 曲の楽譜表現 (SMF) に対し、シンセサイザーを用いて合成した音響信号を用意する。この音響信号を用いてスコアアライメントを行った結果の拍位置と、SMF から算出される拍位置の絶対誤差のパーセンタイルを評価基準として用いる。このような評価方法は、タイミング情報が正確に取れるというメリットがある。また、実際に人間が演奏した録音でも同じような性能を発揮することが示唆されている[7]。

結果を表 1 に示す。人間の拍位置指定精度がおおよそ 100 ミリ秒であることを踏まえると、オーケストラのような複雑な楽器構成をもち音符が密である楽曲でも、人間の拍位置精度と同程度の性能を 7 割方発揮する。HL と MAHL を比較すると、タイミングモデルの有効性が示唆される。MAP-MAHL の結果から、音色と音量を固定した場合は、スペクトルをモデル化するアライメント手法は破綻することが分かる。これは、音色と音量に多様性を持たせることの重要性を表している。

Vol.2011-MUS-91 No.19
2011/7/29

表 1 絶対推定誤差のパーセンタイル [ミリ秒]。小さいほど高精度な推定。HL は時間長を独立に扱った本手法 ($p(T_d) = \delta(T_d - 10)$ に設定) 、MAP-MAHL は音量と音色を固定した本手法、MAHL は提案手法。

		25%	50%	75%	90%	95%
歌声＋	HL	13	37	184	658	1023
ピアノ	MAP-MAHL	749	2175	4811	9973	13737
伴奏	MAHL	7	19	51	119	220
楽器＋	HL	14	32	86	255	473
ピアノ	MAP-MAHL	863	2549	6437	9373	11219
伴奏	MAHL	8	21	45	93	163
ピアノ	HL	17	48	224	891	2040
ソロ	MAP-MAHL	1485	4520	10468	19415	26728
	MAHL	9	21	50	126	269
小規模	HL	16	46	131	393	816
アンサ	MAP-MAHL	1927	4296	8827	16260	25178
ンブル	MAHL	10	22	45	88	133

MAHL の本来の目的である演奏パート間のタイミングの揺らぎの分析の評価は困難である。今回は、MAHL において、個別のパートのアライメントを行ったスペクトログラムを、図示する。**図 2** に示す、三つの和音進行から構成される楽曲フレーズに対してアライメントを行った。演奏パートの割り振りは、和音内の音符がそれぞれ別々のパートに割り当てられるよう適当にヴォイシングを行った。結果を**図 3** に示す。コードのロールが適切に推定されていることが、三つ目のコードのアライメントから推測できる。一方で、最低音の遷移を見ると、一音目から二音目に移り変わる時間が遅れていることがわかる。これは、一音目が二音目がなってからも持続していたためだと考えられる。楽譜に記載されてある音長と違い、実際の演奏はサスティーンペダルといったものにより、楽譜に記載されてあるものと大きく前後する可能性があることに起因する誤推定である。

5. おわりに

本稿では、個別の楽器パートに対するスコアアライメントを算出する手法 MAHL を提案した。評価実験の結果、スコアアライメントとしての有用性が示唆され、また、個別のパートに対するアライメント手法としての有用性が示唆された。今後の課題としては、MAHL のパートアライメント手法としての更なる検証、オンセット検出の統合、音長に対するロバストネスの実現などが考えられる。

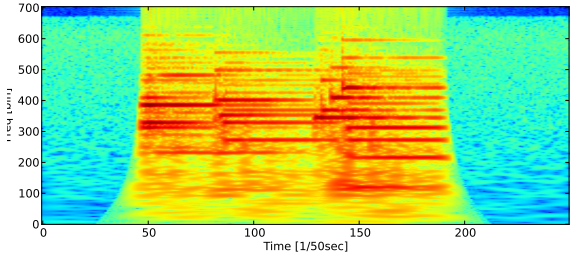

図 2 三つの和音系列から構成される楽曲音の定 Q 変換。

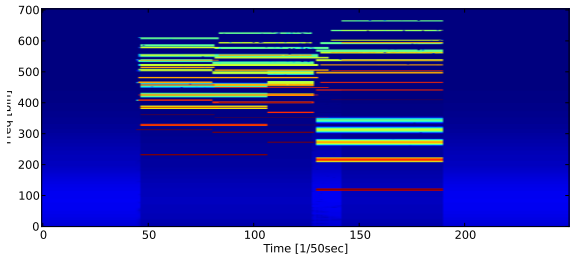

図 3 推定された定 Q 変換。同一和音内における音符の発音時刻のばらつきが確認できる。

参 考 文 献

1) C.Raphael. A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores. In *ISMIR*, pp. 387–394, 2004.
2) N.Hu, et al. Polyphonic audio matching and alignment for music retrieval. In *WASPAA*, pp. 185–188, 2003.
3) A.T. Peeling, P.Cemgil and S.Godsill. A Probabilistic Framework for Matching Music Representations. In *ISMIR*, pp. 267–272, 2007.
4) Sebastian Ewert, etal. High resolution audio synchronization using chroma onset features. In *ICASSP*, pp. 1869–1872, Taipei, Taiwan, April 2009.
5) 吉井和佳, 後藤真孝. 多重音基本周波数解析のための無限潜在的調波配分法. 情報処理学会 第 86 回音楽情報科学研究会, 2010.
6) 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. 日本音響学会 2002 年秋季研究発表会, 第 1-1-7 巻, pp. 649–650, 2002.
7) M.Muller and S.Ewert. Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE TASLP*, Vol.18, No.3, pp. 649–662, March 2010.