拍長の連続性を考慮した潜在的調波配分法に基づくスコアアライメント手法*

☆前澤陽 (京大), 後藤真孝 (産総研), 尾形哲也, 奥乃博 (京大)

近年、計算機を用いて楽譜表現を援用した音楽音響信号の新たな楽しみ方が提唱されている。例えば、過去のヴァイオリン名演奏の指使いを推定する「演奏法の耳コピ」[1] や、自分の嗜好に合致した演奏者の検索 [2] や、特定の楽器を増幅し [3, 4]、市販の CD から、カラオケ音源を作成するといったことが可能となってきている。これらのアプリケーションでは、音楽音響信号の分析のために楽譜情報を用いる。楽譜というシンボリックな情報と音響信号という波形情報の橋渡しするためには、音響信号の位置と楽譜の位置の時間的対応付け (スコアアライメント、以下アライメント) を求めることが必須である。

アライメントに必要な要件は、音色や音量の変化に対するロバストネスと、音符の時系列の適切なモデル化である。特にクラシック音楽では、繰り返しの省略を検出する機構が必要である。というのは、クラシック音楽の楽譜に記載されている繰り返し指示は、しばし演奏者の解釈により、無視されることがあるためである。

従来、音量と音色のロバストネスを実現するため、アドホックな特徴設計 [5–7] や楽器音データベースを用いた音色の学習 [8] を行っていた。しかし、前者には、緻密なパラメータチューニングが必要であり、設計者のチューニングや音源の選定に性能が依存する問題がある。後者には、アライメントの品質が、楽器音データベースの良し悪しに関連する問題がある。また、楽譜の時系列モデル化には、隠れマルコフモデルや、線形動的システム (LDS) がある。前者は、繰り返し構造といった、楽譜上の状態遷移をうまく記述できる。しかし、モデルに暗黙に仮定される音長の独立性は、音楽的に妥当ではなく、これに起因する精度低下が問題となる。後者は、拍の連続性を考慮しているので、このような問題は起こりづらい。しかし、繰り返し構造のような離れた楽譜位置への遷移が扱えないという問題がある。

本稿では、音源の選定が不要であり、かつ音量と音色のロバストネスを実現し、繰り返し構造などを許容し、かつ拍長の連続性を保つアライメント手法を提案する。楽譜時系列のモデル化には、隠れセミマルコフモデル (HSMM) を、LDS による拍長モデルに条件づける。これにより、連続的なテンポと複雑な楽譜構造に対する許容を同時に実現する。音のモデルには、楽器音の混合音スペクトルをベイズ的に扱う音源モデル LHA[9] を用いる。音色と音量に無情報事前分布を置くことにより、これらに対するロバストネスを実現する。また、音色が無情報であるため、音源の選定が不要である。

1 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、特定の楽器が奏でている特有の音高の対を**楽器音高ペア**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。

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。$Z_i^{(i)}(f,d)$ を、状態 d において楽器音高ペア 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、$Z_j^{(h)}(f,i)$ を、周波数 f が、楽器音高ペア i の第 j 倍音から生成される場合 1 の二値行列とする。$Z_{l,d}^{(s)}(t)$ は時刻 t が、状態 d で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。i 番目の楽器音高ペアの基本周波数が μ_i であり、窓関数の影響などにより分散 $\lambda_i^{-1/2}$ 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:

$$p(X|Z^{(i,h,s)},\mu,\lambda) = \prod_{t,i,j,f,d,l} \mathcal{N}\left(\log f/j|\mu_i,{\lambda_i}^{-1}\right)^{Z_{l,d}^{(s)}(t)X(f,t)Z_i^{(i)}(f,d)Z_j^{(h)}(f,i)} \quad (1)$$

調波構造と音量バランスは多項分布に従うと仮定する。

$$p(Z^{(i)}|E,Z^{(s)}) = \prod_{t,i,f,d,l} e_i(d)^{Z_{l,d}^{(s)}(t)X(f,t)Z_i^{(i)}(f,d)} \quad (2)$$

$$p(Z^{(h)}|A,Z^{(i,s)}) = \prod_{t,i,j,f,d,l} a_j(i)^{Z_{l,d}^{(s)}(t)X(f,t)Z_i^{(i)}(f,d)Z_j^{(h)}(f,i)} \quad (3)$$

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:

$$p(\mu,\lambda|\nu,b,m,l) = \prod_i \mathcal{NG}(\mu_i,\lambda_i|m_i^{(H)},b_i^{(H)},l_i^{(H)},\nu_i^{(H)}) \quad (4)$$

$$p(E|E_0) = \prod_d^D \mathrm{Dir}(e(d)|e_0(d)) \quad (5)$$

$$p(A|A_0) = \prod_i^I \mathrm{Dir}(a(i)|a_0(i)) \quad (6)$$

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。

$$p(Z^{(s)}|T,\pi,\tau) = \pi^{Z^{(s)}(1)} \prod_{t=2,l,d,d'\neq d} \left(\tau_{d'}(d)\mathcal{N}\left(\log\frac{l}{\mathcal{L}_d}|T_d,\sigma_T^2\right)\right)^{Z_{1,d'}^{(s)}(t-1)Z_{l,d}^{(s)}(t)} \quad (7)$$

$$p(\pi|\pi_0) = \mathrm{Dir}(\pi|\pi_0) \quad (8)$$

$$p(\tau|\tau_0) = \prod_d \mathrm{Dir}(\tau(d)|\tau_0(d)) \quad (9)$$

---

*Audio-to-Score alignment based on Latent Harmonic Allocation with smoothness of beat length. by Akira MAEZAWA (Kyoto U.), Masataka GOTO (AIST), Tetsuya OGATA (Kyoto U.), Hiroshi G. OKUNO (Kyoto U.)

式 (7) 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:

$$p(T) = \prod_d \mathcal{N}\left(T_d | T_{d-1}, \mathcal{L}_{d-1}{\lambda^{(T)}}_d^{-1}\right) \quad (10)$$

$$p(\lambda^{(T)}) = \prod_d \mathcal{G}({\lambda^{(T)}}_d | l_d^{(T)}, \nu_d^{(T)}) \quad (11)$$

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 $q(LDS, LHA, HSMM)$ が $q_{LDS}(LDS)q_{LHA}(LHA)q_{HSMM}(HSMM)$ 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、$q_{HSMM}(HSMM) = q_{Z^{(s)}}(Z^{(s)})q_\pi(\pi)q_\tau(\tau)$ 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、$q_{LDS}(LDS) = q_T(T)q_{\lambda^{(T)}}(\lambda^{(T)})$ 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。

$$q_Z(Z) \propto \exp\langle \log p(X, LDS, LHA, HSMM)\rangle_{\neg Z} \quad (12)$$

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。

2 モデルの推論

簡単のため、次の変数を定義する:

$$l\mathcal{N}_f(i,j) = \left\langle \log \mathcal{N}\left(\log f/j | \mu_i, {\lambda_i}^{-1}\right)\right\rangle \quad (13)$$
$$= -\frac{1}{2}\left(\frac{\bar{l}_i}{\bar{\nu}_i}(\log f/j - \bar{m}_i)^2 + \frac{1}{\bar{b}_i}\right) - \log 2\pi\bar{\nu}_i + \psi(\bar{l}_i)$$

$$el(d,l) = \langle \log p(\log l | T_d)\rangle_T \quad (14)$$

$$\eta_d(t) = \sum_l \left\langle Z_{d,l}^{(s)}(t)\right\rangle_{Z^{(s)}} \quad (15)$$

$$lA_j(i) = \langle \log a_j(i)\rangle_{a(i)}$$
$$= \psi(\bar{\alpha}_j(i)) - \psi\left(\sum_{l=1}^M \bar{\alpha}_l(i)\right) \quad (16)$$

$$lE_i(d) = \langle \log e_i(d)\rangle_{e(d)}$$
$$= \psi(\bar{\epsilon}_i(d)) - \psi\left(\sum_{l=1}^K \bar{\epsilon}_l(d)\right) \quad (17)$$

$$e\tau'_d(d) = \langle \log \tau_{d'}(d)\rangle_{\tau(d)}$$
$$= \psi(\bar{\tau}'_d(d)) - \psi\left(\sum_{l=1}^D \bar{\tau}_l(d)\right) \quad (18)$$

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。

2.1 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 i が周波数 f に占める割合 $Z^{(i)}$ を次のように更新する:

$$q_{Z^{(i)}}(Z^{(i)}) = \prod_{i,f,d} \gamma_i(f,d)^{Z_i^{(i)}(f,d)} \quad (19)$$

ただし $\gamma_i(f,d) = \frac{\rho_i(f,d)}{\sum_i \rho_i(f,d)}$ であり、ρ は次のように表されるとする:

$$\log \rho_i(f,d) = \left(\sum\nolimits_t X(f,t)\eta_d(t)\right)\times$$
$$\left[lE_i(d) + \sum\nolimits_j \xi_j(f,i)\left(l\mathcal{N}_f(i,j) + lA_j(i)\right)\right] \quad (20)$$

式 (20) 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:

$$q_{Z^{(h)}}(Z^{(h)}) = \prod\nolimits_{j,i,f} \xi_j(f,i)^{Z_j^{(h)}(i,f)} \quad (21)$$

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

$$\log \phi_j(f,i) = \left(\sum\nolimits_{t,d} X(f,t)\eta_d(t)\gamma_i(f,d)\right)\times$$
$$\left[l\mathcal{N}_f(i,j) + lA_j(i)\right] \quad (22)$$

式 (22) も式 (20) 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。

2.1.1 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 $q_E = \prod_i q_{e_i}$ 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_i の事後分布は次のように求められる:

$$q_{e_i} \sim \mathrm{Dir}(e_i | \bar{\epsilon}_i) \quad (23)$$

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

$$q_{a_i} \sim \mathrm{Dir}(a_i | \bar{\alpha}) \quad (24)$$

$$\bar{\alpha}_i = a_0(i) + \sum_{t,f,d} X(f,t)\gamma_i(f,d)\eta_d(t)\xi(f,i)$$

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 $q_{\mu,\lambda} = \prod_i q_{\mu_i,\lambda_i}$ であり、また正規分布と $\mathcal{NG}$ 分布の共役性により、q_{μ_i,λ_i} は次のパラメータで与えられる $\mathcal{NG}(\bar{m}_i^{(H)}, \bar{b}_i^{(H)}, \bar{l}_i^{(H)}, \bar{\nu}_i^{(H)})$ である:

$$\bar{m}_i^{(H)} := \frac{m_i b_i + N_{\gamma,i}\langle \log(f/j)\rangle_{\psi_i}}{b_i + N_{\gamma,i}} \quad (25)$$

$$\bar{b}_i^{(H)} := b_i + N_{\gamma,i} \quad (26)$$

$$\bar{l}_i^{(H)} := l_i + N_{\gamma,i} \quad (27)$$

$$\bar{\nu}_i^{(H)} := \nu_i + \frac{1}{2}\frac{b_i N_{\gamma,i}}{b_i + N_{\gamma,i}}\Bigg(\left(\langle \log(f/j)\rangle_{\psi(i)} - m_i\right)^2$$
$$+ N_{\gamma,i}^2\left\langle \log(f/j) - \langle \log(f/j)\rangle_{\psi(i)}\right\rangle_{\psi(i)}^2\Bigg) \quad (28)$$

$\psi_i(f,j)$ は次のような多項分布である:

$$\psi_{f,j}(i) = \sum_{d,t} \gamma_i(f,d)\xi_j(f,i)\eta_d(t)X(f,t)/N_{\gamma,i} \quad (29)$$

$$N_{\gamma,i} = \sum_{d,t,f,j} \gamma_i(f,d)\xi_j(f,i)\eta_d(t)X(f,t) \quad (30)$$

2.1.2 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 $\psi_l(d) = \sum_{t=2,d'\neq d} \zeta_{d'}(t,l,d)$ 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:

$$\log q_T(T) = \sum_d \sum_l \psi_l(d) \left\langle \log \mathcal{N}\left(\log \frac{l}{\mathcal{L}_d} | T_d, \sigma_T^2\right)\right\rangle_{\sigma_T^2} + \left\langle \log \mathcal{N}\left(T_d | T_{d-1}, \mathcal{L}_{d-1}{\lambda^{(T)}}_d^{-1}\right)\right\rangle_{{\lambda^{(T)}}_d} \quad (31)$$

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:

$$\begin{aligned}
\alpha_d^{(L)}(T_d) &= p(T_d|\psi(1:d)) \\
&\propto \int \alpha_{d-1}^{(L)}(T_{d-1})\, p(T_d|T_{d-1}, \mathcal{L}_{d-1}{\lambda^{(L)}}_d^{-1}) \\
&\quad \times \prod_l p(\log \frac{l}{\mathcal{L}_d}|T_d, \sigma_T^2)^{\psi_l(d)} dT_{d-1} \\
&= \int \mathcal{N}(T_{d-1}|u_{d-1}, s_{d-1}) \mathcal{N}\left(T_d|T_{d-1}, \mathcal{L}_{d-1}{\lambda^{(L)}}_d^{-1}\right) dT_{d-1} \\
&\quad \prod_l \mathcal{N}\left(\log l/\mathcal{L}_d | T_d, \sigma_T^2\right)^{\psi_l(d)} = \mathcal{N}(T_d|u_d, s_d) \quad (32)
\end{aligned}$$

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:

$$s_d^{-1} = \sum_l \psi_l(d) \frac{1}{\sigma_T^2} + \frac{\left\langle {\lambda^{(L)}}_d\right\rangle}{\mathcal{L}_{d-1}} - m_d \left(\frac{\left\langle {\lambda^{(L)}}_d\right\rangle}{\mathcal{L}_{d-1}}\right)^2 \quad (33)$$

$$u_d = s_d \left(m_d \frac{\left\langle {\lambda^{(L)}}_d\right\rangle}{\mathcal{L}_{d-1}} \frac{u_{d-1}}{s_{d-1}} + \sum_l \frac{\psi_l(d)}{\sigma_T^2} \log \frac{l}{\mathcal{L}_d} \right) \quad (34)$$

同様に後ろ向き変数を次のように求める:

$$\begin{aligned}
\beta_d^{(L)}(T_d) &= p(\psi(d+1:T)|T_d) \\
&\propto \int p(\psi(d+2:T)|T_{d+1}) p(T_{d+1}|T_d) \\
&\quad \times \prod_l p(\log \frac{l}{\mathcal{L}_d}, T_{d+1})^{\psi_l(i+1)} dT_{d+1} \\
&= \int \beta_{d+1}(T_{d+1}) \prod_l \mathcal{N}\left(\log \frac{l}{\mathcal{L}_{d+1}} | T_{d+1}, \sigma_T^2\right)^{\psi_l(d+1)} \times \\
&\quad \mathcal{N}\left(T_{d+1}|T_d, \mathcal{L}_d{\lambda^{(L)}}_{d+1}^{-1}\right) dT_{d+1} = \mathcal{N}(T_d|v_d, q_d) \quad (35)
\end{aligned}$$

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:

$$q_d^{-1} = \frac{\left\langle {\lambda^{(T)}}_{d+1}\right\rangle}{\mathcal{L}_d} - n_d \left(\frac{\left\langle {\lambda^{(T)}}_{d+1}\right\rangle}{\mathcal{L}_d}\right)^2 \quad (36)$$

$$v_d = n_d q_d \frac{\left\langle {\lambda^{(T)}}_{d+1}\right\rangle}{\mathcal{L}_d} \left(\sum_l \frac{\psi_l(d+1)}{\sigma_T^2} \log \frac{l}{\mathcal{L}_{d+1}} + \frac{v_{d+1}}{q_{d+1}} \right) \quad (37)$$

これらを用いて、拍長の事後分布を次のように得る:

$$q(T_d|l_{1:T}) = \alpha_d^{(L)}(T_d)\, \beta_d^{(L)}(T_d) = \mathcal{N}\left(T_d | \frac{1}{q_d^{-1} + s_d^{-1}} \left(\frac{v_d}{q_d} + \frac{u_d}{s_d}\right), \frac{1}{q_d^{-1} + s_d^{-1}}\right) \quad (38)$$

2.2 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:

$$\begin{aligned}
el(d,l) &= \left\langle -\frac{1}{2\sigma_T^2} \left(\log l/\mathcal{L}_d - T_d\right)^2 - \log\left(2\pi\sigma_T^2\right) \right\rangle_{T_d} \\
&= -\frac{1}{2\sigma_T^2} \left(\log l/\mathcal{L}_d - \left(\frac{1}{q_d^{-1} + s_d^{-1}} \left(\frac{v_d}{q_d} + \frac{u_d}{s_d} \right) \right) \right)^2 \\
&\quad - \frac{1}{2\sigma_T^2 \left(q_d{}^{-1} + s_d{}^{-1}\right)} - \log\left(2\pi\sigma_T^2\right) \quad (39)
\end{aligned}$$

状態遷移確率 τ の期待値は次のように求まる:

$$q_\tau = \prod_d \mathrm{Dir}\left(\tau_{d'}(d) | \bar{\tau}_{d'}(d)\right) \quad (40)$$

ただし、$\bar{\tau}_{d'}(d) = \tau_{0\,d'}(d) + \sum_{t,l} \zeta_{d'}(t,l,d)$

$q_{Z^{(s)}}(Z^{(s)})$ は次のように求まる:

$$\begin{aligned}
\log q(Z^{(s)}) &= \sum_{l,d} Z_{l,d}^{(s)}(1) \left\langle \log \pi \right\rangle_\pi \\
&\quad + \sum_{t=2,d'\neq d} Z_{1,d'}^{(s)}(t-1) Z_{l,d}^{(s)}(t) (e\tau_{d'}(d) + el(d,l)) \\
&\quad + \sum_{t=1} Z_{l,d}^{(s)}(t) X(f,t) \log \kappa_d(f) \quad (41)
\end{aligned}$$

ただし

$$\begin{aligned}
&\log \kappa_d(f) = \gamma_i(f,d) \\
&\times \left(\sum_{i,j} \xi_j(f,i) \left(l N_f(i,j) + l A_j(i) \right) + l E_i(d) \right) \quad (42)
\end{aligned}$$

$\kappa_d(f)$ 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 $\kappa_d(f)$ 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:

$$\begin{aligned}
&\alpha_{l,d}^{(H)}(t) = p(Z_{(l,d)}^{(s)}(t) = 1 | X(1) \cdots X(t)) \\
&\propto \sum_{l',d'} \alpha_{l',d'}^{(H)}(t) \exp\left\langle \log \tau_{l',d'}(l,d) \right\rangle_\tau \prod_f \kappa_d(f)^{X(f,t)} \\
&\quad = \left(\prod_f \kappa_d(f)^{X(f,t)} \right) \times \left(\alpha_{l-1,d}^{(H)}(t-1) \right. \\
&\quad \left. + \sum_{d'} \exp\left(e\tau_d(d') + el(d,l)\right) \alpha_{1,d'}^{(H)}(t-1) \right) \quad (43)
\end{aligned}$$

$$\begin{aligned}
&\beta_{l,d}^{(H)}(t) = p(X_{t+1}(f) \cdots X_T(f) | Z_{(l,d)}^{(s)}(t) = 1) \\
&= \sum_{l',d'} \beta_{l',d'}^{(H)}(t+1)\, e^{\left\langle \log \tau_{l,d}(l',d') \right\rangle_\tau} \prod_f \kappa_{d'}(f)^{X(f,t+1)} \\
&= \begin{cases}
\left(\prod_f \kappa_d(f)^{X(f,t+1)} \right) \beta_{l-1,d}^{(H)}(t+1) & l > 1 \\
\sum_{d'} \left(\prod_f \kappa_{d'}(f)^{X(f,t+1)} \right) \exp\left(e\tau_{d'}(d)\right) & \\
\times \sum_{l'} \beta_{l',d'}^{(H)}(t+1) \exp\left(el(d',l')\right) & l = 1
\end{cases} \quad (44)
\end{aligned}$$

これらを用い、次の期待値を求める:

$$\eta_d(t) \propto \sum_l \alpha_t^{(H)}(l,d)\,\beta_t^{(H)}(l,d) \qquad (45)$$

$$\xi_{d,l}(d',t) \propto \alpha_{t-1}(1,d')e^{e\tau_{d'}(d)+el(d)}\beta_t(l,d) \qquad (46)$$

3 評価実験

実験では、(1) 現状で多用されているシステムとの性能差 (2)LDS を用いた拍長モデルの有用性、(3) 音色と音量に不確定性を持たせる LHA を用いることの有用性、の三点を評価する。(1) は、クロマベクトルの総コサイン距離最小化基準に基づく DTW を使用する。近年高性能である手法は、クロマベクトル同士の距離を DTW を用いて最小化するものが多い [6]。(2) 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。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 [8] と同等である。(3) 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。スペクトルモデルは [5] と同等になる。調波構造のモデルは [5] 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。

まず、RWC クラシック音楽データベース [10] 60 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 [6]。

結果を表 1 に示す[1] 。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。また、現状多く使用されている手法 (Chroma) より、はるかに性能が高いことが分かる。LH と LHL を比較すると、タイミングモデルの有効性が示唆される。MLHL 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。

Table 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。Chr は従来法、LH は時間長を独立に扱った本手法 ($p(T_d) = \delta(T_d - 10)$ に設定) 、MLHL は音量と音色を固定した本手法、LHL は提案手法。

		25%	50%	75%	90%	95%
歌声+ピアノ伴奏	Chr	88	289	831	2566	7319
	LH	13	37	184	658	1023
	MLHL	749	2175	4811	9973	13737
	LHL	7	19	51	119	220
楽器+ピアノ伴奏	Chr	68	182	619	2714	9848
	LH	14	32	86	255	473
	MLHL	863	2549	6437	9373	11219
	LHL	8	21	45	93	163
ピアノソロ	Chr	90	304	1363	6422	11736
	LH	17	48	224	891	2040
	MLHL	1485	4520	10468	19415	26728
	LHL	9	21	50	126	269
小規模アンサンブル	Chr	90	259	891	2804	4710
	LH	16	46	131	393	816
	MLHL	1927	4296	8827	16260	25178
	LHL	10	22	45	88	133
オーケストラ	Chr	123	394	1384	6688	36550
	LH	38	104	574	4793	16768
	MLHL	3111	10463	21788	34275	44847
	LHL	23	51	119	805	2996

4 まとめ

本稿では、音色や音量の不確定性を扱い、演奏のタイミングモデルを取り入れつつも、繰り返し構造といった、楽譜上の遷移を取り扱えるスコアアライメント手法を提案した。また、音色音量モデルとタイミングモデルの有効性と、現状で多用されている手法の性能差を評価し、その有効性を確認した。今後の課題としては、単一パートのアライメントがある。今までの多くのアライメント手法は、楽譜位置と音響信号の対応付けを求めるが、実際には特定のパートが他より速く弾くといったことがある。単一パートのアライメントを、通常のアライメントから算出出来れば、音源分離や演奏分析といった、楽譜を援用した音楽音響信号分析の性能の向上が期待される。また、信号モデルにアタックや打楽器音も取り入れることにより、更なる精度の向上が期待できる。

[1]本手法によるアライメントの推定結果に基づき拍位置を合成したオーディオデモを以下の URL にて公開している：
http://winnie.kuis.kyoto-u.ac.jp/~amaezaw1/alignment_j

参考文献

[1] A. Maezawa et al. Violin Fingering Estimation Based on Violin Pedagogical Fingering Model Constrained by Bowed Sequence Estimation from Audio. In *IEA/AIE*, 2010.

[2] A. Maezawa et al. Query-By-Conducting: An Interface to retrieve classical-music interpretations by real-time tempo input. In *ISMIR*, pages 477–482, 2010.

[3] K. Itoyama et al. Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals. In *ICASSP*, pages I–57–I–60, April 2007.

[4] Y. Han and C. Raphael. Desoloing Monaural Audio Using Mixture Models. In *ISMIR*, pages 145–148, 2007.

[5] C. Raphael. A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores. In *ISMIR*, pages 387–394, 2004.

[6] M. Muller and S. Ewert. Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE TASLP*, 18(3):649–662, March 2010.

[7] N. Hu et al. Polyphonic audio matching and alignment for music retrieval. In *WASPAA*, pages 185–188, 2003.

[8] A.T. Peeling, P. Cemgil and S. Godsill. A Probabilistic Framework for Matching Music Representations. In *ISMIR*, pages 267–272, 2007.

[9] K. Yoshii and M. Goto. Infinite Latent Harmonic Allocation: A nonparametric Bayesian approach to multipitch analysis. In *ISMIR*, pages 309–314, 2010.

[10] M. Goto. Development of the RWC Music Database. In *Int'l Congress on Acoustics*, volume I, pages 553–556, 2004.