5R-4

潜在的調波配分法に基づく隠れセミマルコフモデルを用いたベイズ的スコアアライメント

前澤 陽 † **後藤 真孝** ‡ **尾形 哲也** † **奥乃 博** †
† 京都大学大学院 情報学研究科 知能情報学専攻 ‡ 産業技術総合研究所 (AIST)

1. はじめに

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, 6, 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。手法の概念図を図 1 に示す。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。

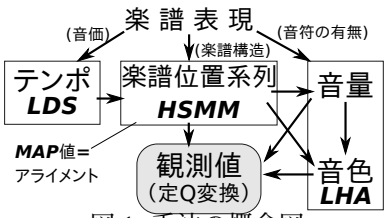

図 1: 手法の概念図

Bayesian score alignment based on Latent Harmonic Allocation using Hidden Semi-Markov Model. *Akira Maezawa (Kyoto U.), Masataka Goto (AIST), Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto U.)*

2. 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。音色と音量のロバストネスを実現するために、これらを固定せず、変化を許容するようなモデル化を行う。また、複雑な楽譜構造に対処するため、楽譜を HSMM として、モデル化する。また、音楽的に妥当な音長のモデルを実現するために、楽譜時系列の事前分布に滑らかな音長を保つようなモデルを設定する。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。すると、これらを更に確率変数としてして扱い、特定の値に固着させないよう (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。

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

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

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

式 (7) は、楽譜時系列を、拍の長さ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 d 番目の音符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 $\arg\max\sum_l Z^{(s)}_{l,d}(t)$ 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 において、事後分布において、μ と λ を除く、すべての確率変数の独立性を仮定する。紙面の制約上、導出は省略するが、混合ガウス分布の推論と、HSMM の前向き後ろ向きアルゴリズムと、カルマンスムーザーを組み合わせた物に類似した更新式が導出できる。

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 Dynamic Time Warping (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。これには、音価に比例するような音長の期待値を持った HSMM を使用した。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、楽譜表現に記載されてい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。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 割方以上発揮する。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。ML-LHL 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。

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

参考文献

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

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

[3] Itoyama, K., et al.: Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals, in *ICASSP*, pp. I–57–60 (2007).

[4] Han, Y. and Raphael, C.: Desoloing Monaural Audio Using Mixture Models, in *ISMIR*, pp. 145–148 (2007).

[5] Raphael, C.: A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores, in *ISMIR*, pp. 387–394 (2004).

[6] Muller, M. and Ewert, S.: Towards Timbre-Invariant Audio Features for Harmony-Based Music, *IEEE TASLP*, Vol. 18, No. 3, pp. 649–662 (2010).

[7] Hu, N., et al.: Polyphonic audio matching and alignment for music retrieval, in *WASPAA*, pp. 185–188 (2003).

[8] Peeling, A., P. Cemgil and Godsill, S.: A Probabilistic Framework for Matching Music Representations, in *ISMIR*, pp. 267–272 (2007).

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

[10] Goto, M.: Development of the RWC Music Database, in *Int'l Congress on Acoustics*, pp. 553–556 (2004).