Vol.2012-MUS-96 No.13
2012/8/10

隠れセミマルコフモデルと線形動的システムを組み合わせた音楽音響信号と楽譜の実時間アライメント手法

山本 龍一[1,a)]　酒向 慎司[1,b)]　北村 正[1,c)]

概要：本稿では，楽譜に基づく音楽音響信号から，演奏位置とテンポを推定する問題について論じる．隠れセミマルコフモデル（HSMM）に基づく演奏位置推定と，線形動的システム（LDS）に基づくテンポ推定を組み合わせることで，入力信号の未来の情報が使えない制約の元で効果を発揮する実時間拍予測アルゴリズムを提案する．具体的には，遅延を許容して信頼性のある演奏位置を推定し，テンポを用いて現在位置を予測する．クラシック音楽およびジャズ音楽データベースを用いてオンセット検出に関する評価実験を行った結果，提案する実時間拍予測アルゴリズムを用いることで，許容誤差 300ms において約 15% 精度が向上することが確認された．

キーワード：音響信号と楽譜の実時間アライメント，楽譜追跡，隠れセミマルコフモデル，線形動的システム

Real-time Audio to Score Alignment Using a Hybrid Hidden Semi-Markov Model and Linear Dynamical System

Ryuichi Yamamoto[1,a)]　Shinji Sako[1,b)]　Tadashi Kitamura[1,c)]

Abstract: In this paper, we propose a real-time audio to score alignment method which jointly estimates performer's beat position and continuous tempo based on a Hidden Semi-Markov Model (HSMM) and Linear Dynamical System (LDS). Also, we propose an effective on-line beat prediction algorithm with the use of two optimal algorithm of HSMM based beat estimation and LDS based tempo estimation which performs well in the real-time setting. More specifically, a reliable beat position is estimated for some delay and then current beat position is predicted using tempo. Experimental results on the audio onset detection task for piano database of classical music and jazz music database showed that our method obtained 15% higher results within 300ms tolerance compared to the method with no use of our beat anticipation algorithm.

Keywords: Real-time Audio to Score Alignment, Score Following, Hidden Semi-Markov Model, Liniear Dynamical System

1. はじめに

本研究では，楽譜に基づく演奏に対して，音楽音響信号から実時間で演奏位置とテンポを推定する問題（以下，実時間アライメントと呼ぶ）を扱う．このような技術は，演奏に合わせて伴奏を自動再生させる自動伴奏，歌唱ロボット，音楽ライブにおける音響信号加工，音楽と映像の同期など様々な応用が考えられるため，非常に有用だといえる．

実時間アライメントの問題の性質として注目すべきなのは，入力される信号に対して瞬時に演奏位置を推定しなければならない点である．そのため，入力信号の未来の情報が使えないという厳しい制約がある．本研究では，この制約の中で効果的に演奏位置とテンポを推定する方法について議論する．

[1] 名古屋工業大学大学
Nagoya Institute of Technology
[a)] ryuichi@mmsp.nitech.ac.jp
[b)] sako@mmsp.nitech.ac.jp
[c)] kitamura@mmsp.nitech.ac.jp

ⓒ 2012 Information Processing Society of Japan

音響信号と楽譜のアライメントの問題には主に二つの困難がある．一つは，音響信号は MIDI などの記号表現とは異なり，音高，発音時刻に関する情報が陽に得られないことである．複数の音の同時発音，残響，その他ノイズなどの影響によって信号は非常に複雑になり，問題を難しくしている．もう一つは，人間の演奏の多様性に起因するものである．楽譜に基づく演奏であっても，演奏誤り，繰り返し構造の無視，テンポの揺らぎなど，人間の演奏には多くの不確実性が存在する．

上述の問題に対して，隠れマルコフモデル（Hidden Markov Model, HMM）や DP（Dynamic Programming）マッチングに基づく方法 [1,2] と，状態空間モデルを用いた方法 [3–5] が提案されている．HMM に基づく手法では，演奏の誤り，繰り返し構造の無視に対応可能である一方で，演奏位置を推定する際に入力信号の系列全体を用いる最適化アルゴリズムが必要となる．これは DP マッチングでも同様である．そのため，逐次的に信号が入力される場合には，必ずしも適した方法であるとは言えない．一方で，状態空間モデルに基づく手法では効果的な拍予測アルゴリズムが提案されている [6]．しかし，それらの手法は演奏位置とテンポを同時に推定できる枠組みを持つが，繰り返し構造の無視は考慮されていない．

そのような背景に基づき，本研究では二つの手法の利点を備えた実時間アライメント手法を提案する．演奏の生成過程を隠れセミマルコフモデル（Hidden Semi-Markov Model, HSMM）でモデル化することで，繰り返し構造の無視に対応可能な演奏位置推定アルゴリズムを導き，テンポの変化を線形動的システム（Linear Dynamical System, LDS）でモデル化することで，演奏位置の予測に基づくテンポの反復推定アルゴリズムを導く．同様の手法として Cont らの手法 [7] があるが，我々は演奏位置とテンポの最適化アルゴリズムを組み合わせた新しい実時間拍予測アルゴリズムを提案する．

以下の構成は次の通りである．第 2 章では HSMM に基づく音響信号と楽譜のアライメントについて述べ，第 3 章では LDS に基づくテンポ推定について述べる．第 4 章では拍予測に基づく実時間アライメントアルゴリズムについて述べ，第 5 章では評価実験について報告する．第 6 章では結論と今後の展望について述べる．

2. HSMM に基づく音響信号と楽譜のアライメント

2.1 問題設定

音響信号と楽譜のアライメント問題は，入力の音響信号に対して楽譜上の演奏位置を求める問題である．しかし，第 1 章で述べた通り，音響信号からは音高，発音時刻に関する情報が陽に得られず，人間の演奏には演奏誤り，繰り返し構造の無視，テンポ変化など多くの不確実性が含まれるため容易ではない．ここでは，そのような演奏に対して，楽譜上の演奏位置を推定する問題について議論する．

以降，演奏位置とは楽譜上で同時発音されるべき音符を一つの集合として，そのインデックスを表すとする．以下，和音，または拍位置と呼ぶ．音響信号はサンプリング周波数によって離散化された離散時間信号列であり，一定のフレームシフトでオーバーラップして解析され，特徴系列が抽出される．以下，特徴系列に対するインデックスをフレーム番号，または単に時刻と呼ぶ．テンポは楽譜上の一拍当たりの実際の演奏時間 [秒/拍] を意味するとする．楽曲の楽譜情報としては，広く普及している MIDI を用いる．ただし，実際には楽譜上の各音符の発音の時間（楽譜上の拍の位置を示す）と音高の情報のみを用いる．

2.2 問題の定式化

入力の音響信号から抽出される特徴系列を $\mathbf{O} = \{\boldsymbol{o}_t\}_{t=1}^{T}$，$\boldsymbol{o}_t \in \mathbb{R}^n$，$T$ は総フレーム数とする．一和音に相当するフレーム群を一つのセグメントとして，総セグメント数を N，セグメントの系列（以下，セグメンテーションと呼ぶ）を $\mathbf{Q} = \{q_n\}_{n=1}^{N}$，$q_n = (t_n^s, t_n^e, s_n)$ とする．t_n^s は n 番目のセグメントの先頭，t_n^e は最後尾，s_n はセグメントに対応する拍位置を示す．ただし，t_n を n 番目のセグメントの最後尾とし，$t_n^s = t_n - d_n + 1$，$t_n^e = t_n$ である．セグメント内の部分特徴系列は $\boldsymbol{o}_{t_n-d_i+1:t_n} = \{\boldsymbol{o}_t\}_{t=t_n-d_n+1}^{t_n}$ のように表す．楽譜上の拍位置の系列を $\mathbf{S} = \{s_n\}_{n=1}^{N}$，$s_n \in \{1, \dots M\}$，対応するセグメント長（以下，状態継続長と呼ぶ）を $\mathbf{D} = \{d_n\}_{n=1}^{N}$，$d_n \in \{1, \dots D\}$，テンポを $\mathbf{R} = \{r_n\}_{n=1}^{N}$，$r_n \in \mathbb{R}$ とする．M は総和音数，D は最大状態継続長である．$\sum_{n=1}^{N} d_n = T$ であり，状態継続長の総和と総フレーム数は等しいことに注意する．

以上を踏まえて，音響信号と楽譜のアライメントの問題を次のように定式化する．

$$\hat{\mathbf{Q}} = \arg\max_{\mathbf{Q}\in\mathcal{Q}} p(\mathbf{Q}|\mathbf{O}) = \arg\max_{\mathbf{Q}\in\mathcal{Q}} p(\mathbf{Q}, \mathbf{O}) \qquad (1)$$

$\mathcal{Q}$ は考えうるあらゆるセグメンテーションの集合とする．

2.3 HSMM による演奏のモデル化

楽譜に基づく演奏は，楽譜上の拍位置を一つの状態とした HSMM として捉えられる．HSMM によるモデル化の概念を**図 1** に示す．HMM では観測信号の独立性の仮定のため，発音される音長に関する制約を直接モデル化することができなかった，一方で，HSMM では直接モデル化することができる．HSMM についての詳細は，文献 [8] を参照する．

ここで，各状態は一つ前の状態のみに依存すると仮定すれば，式 (1) における同時確率は以下のように与えられる．

ⓒ 2012 Information Processing Society of Japan

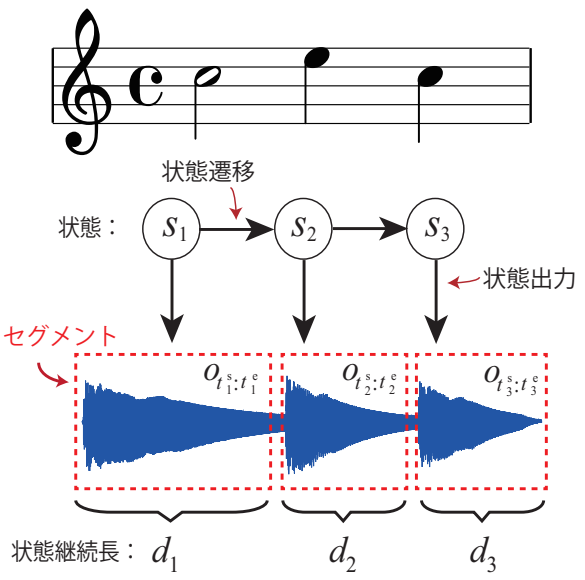

図 1 HSMM に基づく音響信号のモデル化の概念.
Fig. 1 The modeling concept of audio signal based on HSMMs.

$$p(\mathbf{Q},\mathbf{O}) = p(s_1)p(d_1|s_1,r_1)p(o_{t_1^s:t_1^e}|s_1) \cdot \prod_{n=2}^{N} p(s_n|s_{n-1})p(d_n|s_n,r_n)p(o_{t_n^s:t_n^e}|s_n) \quad (2)$$

ただし，$p(s_n|s_{n-1})$ は状態遷移確率，$p(d_n|s_n,r_n)$ は状態継続長に関する確率，$p(o_{t_n^s:t_n^e}|s_n)$ は状態出力確率である．それらによって，HSMM は記述される．状態継続長に関する確率はテンポに依存しているが，Viterbi アルゴリズムによる推論を可能にするため，ここではテンポは固定する．

2.4 観測モデル

2.4.1 特徴量

既存研究において，音響信号と楽譜のアライメントにおいてどのような音響特徴が有効かどうか比較分析が行われている [9]．その研究によれば，対数周波数領域の振幅スペクトルが有効であると示されている．その他にクロマベクトルも考えられるが，予備実験で比較した結果，本研究では連続ウェーブレット変換によって得られる振幅スペクトル（以下，スペクトルと呼ぶ）を用いる．

2.4.2 楽譜からのスペクトル生成モデル

楽譜は演奏の記号表現であるため，そのままでは音響信号と比較することができない．本研究では，ガウス混合モデルに基づく楽譜からのスペクトル生成モデルを考え，スペクトル同士の距離を観測モデルとして用いる．

拍位置 $s_n = i$ に指定されている音符に対応する基本周波数の集合を $\boldsymbol{p}_i = \{p_{i,h}\}_{h=1}^{H_i}$ として，スペクトル $\boldsymbol{x}_i = \{x_{i,f}\}_f$ を以下のようにモデル化する．

$$x_{i,f} = (1-q)\sum_{h=1}^{H_i}\sum_{k=1}^{K} v_{h,k}\mathcal{N}(f;\log(k\cdot p_{i,h}),\sigma^2) + qU(f) \quad (3)$$

K は整数倍の倍音数，$v_{h,k}$ はガウス分布に対する重み，σ^2 はガウス分布の分散，$U(f)$ は一様分布，q は一様分布に対する重みを表す．$f = \log f'$ とし，周波数は対数であることに注意する．実験では，$K = 5$，$v_{h,k} \sim 1/2^k$，$\sigma = 70$ [cent]，$q = 0.01$ とした．

2.4.3 スペクトルに対する距離尺度

スペクトル間の距離には，カルバック・ライブラー情報量（Kullback-Leibler divergence）を用いる．

拍位置 $s_n = i$ におけるスペクトル $\boldsymbol{o}_t$ に対する状態出力確率を以下のようにモデル化する．

$$p(\boldsymbol{o}_t|s_n = i) = \exp\Big(-g\cdot D^{\mathrm{KL}}(\boldsymbol{o}_t||\boldsymbol{x}_i)\Big) \quad (4)$$

g は指数関数の減衰速度を制御するパラメータである．セグメント q_n における部分特徴系列 $o_{t_i^s:t_i^e}$ に対する状態出力確率は，セグメント内の観測信号の独立性を仮定して，以下のようにモデル化する．

$$p(\boldsymbol{o}_{t_n^s:t_n^e}|s_n = i) = \prod_{t=t_n^b}^{t_n^e} p(\boldsymbol{o}_t|s_n = i) \quad (5)$$

2.5 状態遷移モデル

楽譜に基づく演奏では，通常は楽譜通りに正しく拍位置が遷移していくと考えられるが，音抜け，繰り返し構造の無視を考慮する必要がある．HSMM においては，状態間の任意の遷移に対して適切な遷移確率を与えることによって，そのような問題に対処できる．

本研究では，拍位置 $s_{n-1} = i$ から $s_n = j$ に遷移する確率を以下のようにモデル化する．

$$p(s_n = j|s_{n-1} = i) = \begin{cases} p^{next} & ,j = i+1 \\ 0 & ,j = i \\ (1-p^{next})/(M-2) & ,\text{otherwise} \end{cases} \quad (6)$$

p^{next} は，次の拍位置への遷移確率であり，十分大きな確率とする．状態継続長は別にモデル化されるため，自己遷移はしないとする．それ以外の遷移は非常に低い確率になると考えられるので，ここでは簡単のため等確率とした．繰り返し構造が無視されやすい箇所が限定されるなど，演奏に傾向が見られる場合には，その傾向に応じて経験的に確率を与えることができる．

2.6 状態継続長モデル

演奏における発音時間は，拍の長さとテンポを掛け合わせたものになると予想される．

拍位置 $s_n = i$ における状態継続長 $d_n = d$ に対する確率は，拍の長さを l_n として，以下のガウス分布でモデル化する．

$$p(d_n = d|s_n = i) = \mathcal{N}(d;r_n l_n,\sigma_i^2) \quad (7)$$

ⓒ 2012 Information Processing Society of Japan

2.7 Viterbi アルゴリズムによる拍位置の推定

式 (1) を満たす HSMM における最尤状態系列は，HMM の場合と同様に Viterbi アルゴリズムを用いて求めることができる．

与えられた特徴系列 $\mathbf{O} = \{\boldsymbol{o}_t\}_{t=1}^{T}$ に対して，時刻 t において拍位置 $s_t = i$ である確率（以下，前向き確率と呼ぶ）を以下のように定義する．

$$\begin{aligned}\alpha_t(i) &= \max_{s_1,\ldots,s_t}\Big[p(s_1,\cdots,s_t,\boldsymbol{o}_{1:t})\Big] \\ &= \max_{j,d=1,\ldots,D}\Big[b_i(\boldsymbol{o}_{t-d+1:t})p_i(d)a_{i,j}\alpha_{t-d}(j)\Big] \quad (8)\end{aligned}$$

ただし，i, j は状態に関するインデックス，d は状態継続長に関するインデックスある．表記の簡単のため，$a_{i,j}$ は状態遷移確率，$b_i(\boldsymbol{o}_{t-d+1:t})$ は状態出力確率，$p_i(d)$ は状態継続長に関する確率とした．また，$\alpha_0(i) = \pi_i$ は初期状態確率とする．

上記の前向き確率を用いると，以下のようにして最尤状態系列を求めることができる．

$$(\hat{s}_n, \hat{d}_n) = \arg\max_{i,d}\Big[\alpha_{t_n^e}(i,d)\Big] \quad (9)$$

$$t_{n-1}^e = t_n^e - \hat{d}_n \quad (10)$$

説明の都合上，前向き確率において d に関して最大化を行わない変数を以下のように定義した．

$$\alpha_t(i,d) = \max_j\Big[b_i(\boldsymbol{o}_{t-d+1:t})p_i(d)a_{i,j}\alpha_{t-d}(j)\Big] \quad (11)$$

実際には，前向き確率の計算の段階で状態継続長に関して確率は最大化されることに注意する．

以上より，Viterbi アルゴリズムによる拍位置の推定アルゴリズムを以下に示す．

Step 1 式（8）を $t = 1,\ldots,T$，$i = 1,\ldots,M$ に対して計算する．

Step 2 $t_N^e = T$ として，式（9）と式（10）を $n = N, N-1,\ldots$ に対して $t_{n-1}^e = 0$ となるまで繰り返す．

3. LDS に基づくテンポ推定

3.1 問題設定

演奏におけるテンポは，連続する音符の発音時間間隔（Inner Onset Interval, IOI）によって知覚されるものと考えられる．しかし，第 1 章で述べた通り，音響信号から発音時刻は陽に得られない．ここでは，HSMM に基づくアライメントによって得られるセグメンテーションの結果（推定された IOI の系列）からテンポを推定する問題を考える．ただし，簡単のため，演奏は楽譜に基づき順方向に正しく進むと仮定する．

ここで考慮すべきなのは，テンポは時間的に変化する性質を持つことと，セグメンテーションの結果には推定誤差が含まれることである．したがって，テンポ推定の方法は誤差を考慮しつつ，かつテンポの時間変化をモデル化できることが望ましい．ここではテンポは時間的に滑らかに変化すると仮定して，そのような性質を持つテンポを推定する問題について議論する．

3.2 問題の定式化

HSMM に基づくアライメントによって得られるセグメンテーションを $\mathbf{Q} = \{q_n\}_{n=1}^{N}$，IOI の系列を $\mathbf{D} = \{d_n\}_{n=1}^{N}$，テンポを $\mathbf{R} = \{r_n\}_{n=1}^{N}$ とする．IOI に対応する楽譜上の拍の長さを $\mathbf{L} = \{l_n\}_{n=1}^{N}$ とする．ここで，IOI は HSMM における状態継続長に等しいことに注意する．

以上を踏まえ，テンポ推定の問題を次のように定式化する．

$$\hat{\mathbf{R}} = \arg\max_{\mathbf{R}\in\mathcal{R}} p(\mathbf{R}|\mathbf{Q}) = \arg\max_{\mathbf{R}\in\mathcal{R}} p(\mathbf{R},\mathbf{Q}) \quad (12)$$

$\mathcal{R}$ は考えうるテンポの集合である．

3.3 LDS によるテンポのモデル化

テンポが局所的にはほぼ一定であると仮定すると，テンポの変化および IOI としての観測の過程が線形で表される LDS でモデルできる．

$$r_n = r_{n-1} + w_n \quad (13)$$

$$d_n = r_n l_n + v_n \quad (14)$$

式（13）はテンポの変化の過程を表し，式（14）は IOI としての観測の過程を表している．誤差 w_n, v_n は平均が 0，分散 U, V の独立なガウス分布に従うとする．

また，テンポに関するマルコフ性の仮定から，式（12）における同時確率は以下のように与えられる．

$$\begin{aligned}p(\mathbf{R},\mathbf{Q}) &= p(r_1)p(d_1|r_1,s_1) \\ &\quad\cdot\prod_{n=2}^{N} p(r_n|r_{n-1})p(d_n|r_n,s_n) \quad (15)\end{aligned}$$

ただし，$p(r_n|r_{n-1})$ は状態遷移確率，$p(d_n|r_n)$ は状態出力確率であり，それぞれ以下のガウス分布に従う．s_n はセグメンテーションの結果として与えられることに注意する．

$$p(r_n|r_{n-1}) = \mathcal{N}(r_n; r_{n-1}, U) \quad (16)$$

$$p(d_n|r_n,s_n) = \mathcal{N}(d_n; r_n l_n, V) \quad (17)$$

3.4 カルマンフィルタによるテンポ推定

式（12）を満たすテンポは，カルマンフィルタと呼ばれるアルゴリズムによって推定することができる．カルマンフィルタの詳細については，文献 [10] を参照する．

以下に 3.3 節で述べたモデルに対するカルマンフィルタによるテンポ推定のアルゴリズムを示す．テンポの推定値を μ_n，テンポの誤差の分散を P_n，カルマンゲインを K_n とした．

ⓒ 2012 Information Processing Society of Japan

Step 1 $n = 0$ とする．初期値 $\hat{\mu}_0 = \mu_0$， $\hat{P}_0 = P_0$ を設定する．

Step 2 テンポを予測する．

$$\hat{\mu}_{n+1} = \mu_n \tag{18}$$
$$\hat{P}_{n+1} = U + P_n \tag{19}$$

Step 3 観測 d_{n+1} を用いてテンポを更新する．

$$K_{n+1} = \hat{P}_{n+1} l_{n+1} / (l_{n+1}^2 \hat{P}_{n+1} + V) \tag{20}$$
$$\Sigma_{n+1} = (1 - K_{n+1} l_{n+1}) \hat{P}_{n+1} \tag{21}$$
$$\mu_{n+1} = \hat{\mu}_{n+1} + K_{n+1}(d_{n+1} - \hat{\mu}_{n+1} l_{n+1}) \tag{22}$$

Step 4 $n = n + 1$ として，Step 2 に戻る．

4. 拍予測に基づく実時間アライメントアルゴリズム

4.1 Viterbi アルゴリズムのオンライン近似の問題点

2.7 節で述べた Viterbi アルゴリズムを実時間アライメントに用いる際の問題点は，拍位置の最尤推定値が必ずしも正しいとは限らないことである．Cont らは，時刻 t における拍位置を以下のように Viterbi アルゴリズムをオンライン近似することによって求めている [7].

$$\hat{s}_t = \arg\max_i [\alpha_t(i)] \tag{23}$$

しかし．逐次的に音響信号が入力される状況では，時刻 t における最適解が，時刻 $t + t'$ においては最適解ではないといったことが起こりうる．特に，音響信号が複雑な場合にはその傾向が顕著である．これに対する解決策の一つとして，ある程度の遅延を許容してアライメントを取る方法が考えられるが，実時間アライメントの問題では，遅延は大きなリスクであり，好ましい方法ではない．

このような問題に対して我々は，HSMM に基づく拍位置の推定と，LDS によって推定されたテンポに基づく拍予測を組み合わせたオンラインアルゴリズムを提案する．

4.2 実時間拍予測アルゴリズム

提案法では，HSMM に基づくアライメントにおいて α 秒の遅延を許容することで信頼性のある拍位置を推定し，LDS によるテンポの推定値から現在の時刻における拍位置を予測する．HSMM に基づくアライメントにおいて α 秒の遅延が発生するが，テンポを用いた予測によって補うことができる．また，カルマンフィルタによって拍位置の予測誤差が減少するようにテンポは推定される．

時刻 t における HSMM に基づくアライメントから得られるフレーム単位の拍推定の結果を $\{\hat{s}_1, \ldots, \hat{s}_{t-\alpha}, \ldots, \hat{s}_t\}$，LDS に基づくテンポ推定の結果を，$r_t^{-1} = 1/r_t$ [拍/ 秒] として $\{\hat{r}_1^{-1}, \ldots, \hat{r}_{t-\alpha}^{-1}, \ldots, \hat{r}_t^{-1}\}$ とする．楽譜上の絶対的な拍位置を $\{b_1, \ldots, b_{t-\alpha}, \ldots, b_t\}$ とすると，現在の拍位置の予測値 $\hat{b}_t$ は，以下の予測式により与えられる．

$$\hat{b}_t = b_{t-\alpha} + \int_{t-\alpha}^{t} r_\tau^{-1} d\tau \tag{24}$$

ただし，実際には α は小さい値として，区間内でテンポが一定だと仮定すると，予測式は以下のように表される．

$$\hat{b}_t = b_{t-\alpha} + r_{t-\alpha}^{-1} \cdot \alpha \tag{25}$$

ここで，遅延の幅 α をどのように決めるかという問題がある．α を大きくすれば，HSMM に基づくアライメントの精度は向上すると考えられるが，拍予測による予測誤差が拡大する恐れがある．このように，遅延の幅と拍予測の精度にはトレードオフの関係がある．予備実験では，セグメンテーションの結果に対して一つ前のセグメントまでの時間を遅延幅とすると一番良い結果が得られた．

以上の議論に基き，拍予測に基づく実時間アライメントアルゴリズムを以下に示す．

Step 1 Viterbi アルゴリズムにより入力信号のセグメンテーションを行う．

Step 2 セグメンテーションの結果からカルマンフィルタによってテンポの推定を行う．

Step 3 遅延幅を一つ前のセグメントまでの時間と設定する．

Step 4 予測式に基いて現在の拍位置を予測する．

Step 5 信号が入力される度に，Step 1 に戻って繰り返す．ただし，時間が経過するごとにセグメント数が増加するので，セグメンテーションを $\{q_n\}_{n=n'}^{N}$ として，$N - n'$ が定数となるように n' を定めることで計算量の増大を回避する．

5. 評価実験

5.1 実験条件

提案手法の有効性を検証するため，拍予測アルゴリズムを用いる場合と用いない場合でオンセット検出に関する比較実験を行った．拍予測を用いない場合は，4.3 節で示したアルゴリズムの Step 3 と Step 4 を省略した．実行環境は，MacOS X 64bit，プロセッサは 1.8 GHz Intel Core i7，メモリは 4GB である．提案アルゴリズムは，和音数の 2 乗に比例して計算量が増大するため，実験で用いた環境では和音数が約 300 以上で実時間で動作しなかった．ここでは処理時間については考慮せず，実時間アルゴリズムで処理することによって評価を行った．

楽曲には，多声音楽に対する頑健性とテンポ変化に対する頑健性を検証するために，クラシック音楽のピアノデータベース MAPS [11] から YAMAHA Disklavier によって演奏された 35 曲と，RWC ジャズ音楽データベース [12] から RWC-MDB-J-2001-M01 の 15 曲を用いた．評価基準はオンセット検出の再現率とし，誤差を 50ms，100ms，300ms，500ms の 4 段階の許容範囲を設定した．なお，正

ⓒ 2012 Information Processing Society of Japan

確な正解率を算出するために，音響信号は MIDI データと完全に同期されたものを用いた．RWC の楽曲は MIDI と音響信号にずれが含まれたため，YAMAHA XG WDM SoftSynthesizer を音源として MIDI ファイルから同期された wav ファイルを作成して実験に用いた．音響信号はモノラル，サンプリング周波数 10kHz にダウンサンプリングして用いた．周波数解析にはガボールウェーブレット変換を用い，フレームシフトを 10ms，周波数解像度を 25cent，帯域を 55Hz から 3523Hz とした．実験でのパラメータは表 1 の通りである．拍予測を用いる場合と用いない場合で別々にパラメータの最適化を行った結果，重み g は，拍予測を用いる場合は 0.05，用いない場合は 0.1 とした．

表 1 提案手法におけるパラメータ．
Table 1 parameter settings.

パラメータ	値
状態出力確率に関する重み係数 g	0.05 or 0.1
状態遷移確率 p_{next}	0.95
状態継続長におけるガウス分布の 標準偏差 $\sqrt{\sigma}$	0.4 [秒]
テンポに関する誤差の標準偏差 $\sqrt{U}$	0.3 [秒/拍]
IOI に対する誤差の標準偏差 $\sqrt{V}$	0.7 [秒]

5.2 実験結果

拍予測アルゴリズムを用いない場合を（A），用いる場合を（B）として，MAPS に対する実験結果を**表 2** に，RWC-JAZZ に対する実験結果を**表 3** に示す．両方のデータベースにおいて，提案する拍予測アルゴリズムを用いた場合に大幅に精度が向上することが示された．RWC-JAZZ の方が全体的に精度が高い原因は，実音源ではなくソフトウェア音源によって合成した wav ファイルを用いたためだと考えられる．拍位置の予測を用いない場合には，オンセットの検出遅れ，別の状態への誤推定が見られたが，提案法ではそのような問題が緩和されるのを確認した．

表 2 MAPS 35 曲に対するオンセット検出の結果（%）．
Table 2 Onset detection results on the 35 songs from MAPS database (%).

手法	< 50ms	< 100ms	< 300ms	< 500ms
A	14.5	42.1	73.0	77.7
B	**52.2**	**74.8**	**87.5**	**89.4**

表 3 RWC-JAZZ 15 曲に対するオンセット検出の結果（%）．
Table 3 Onset detection results on the 15 songs from RWC-JAZZ database (%).

手法	< 50ms	< 100ms	< 300ms	< 500ms
A	19.1	48.1	77.2	81.5
B	**70.2**	**81.4**	**92.5**	**95.2**

6. おわりに

本稿では，HSMM に基づく演奏位置の推定と LDS に基づくテンポ推定を組み合わせた音響信号と楽譜の実時間アライメント手法を提案した．演奏位置とテンポの最適化アルゴリズムを組み合わせることで，Viterbi アルゴリズムのオンライン近似による問題点を緩和可能な実時間拍予測アルゴリズムを示した．オンセット検出に関する評価実験では，クラシック音楽とジャズ音楽データベースの両方において，提案する拍予測アルゴリズムを用いることで精度が向上し，多声音楽に対する頑健性とテンポ変化に対する頑健性の向上が達成された．今後の課題としては，繰り返し構造の無視に対する頑健性の評価，拍位置の推定精度と予測誤差のトレードオフを考慮した遅延幅の最適化，アルゴリズムの高速化などを検討している．

謝辞 本研究は，名古屋工業大学研究推進経費の支援を受けた．

参考文献

[1] Dixon, S.: An On-line Time Warping Algorithm for Tracking Musical Performances, *Proc. IJCAI* pp.1727–1728, (2005).
[2] Orio, N., Dechelle, F.: Score Following Using Spectral Analysis and Hidden Markov Models, *Proc. ICMC* pp.151–154, (2001).
[3] Duan, Z. and Pardo, B.: A state space model for online polyphonic audio-score alignment, *Proc. ICASSP* pp.197–200, (2011).
[4] Raphael, C.: Aligning music audio with symbolic scores using a hybrid graphical model, *Machine Learning*, vol.65 no.2-3, pp389–409, (2006).
[5] Otsuka, T., Nakadai, K., Takahashi, T., Ogata, T. and Okuno, H. G.: Real-time audio-to-score alignment using particle filter for coplayer music robots, *EURASIP J. Adv. Signal Process*, (2011).
[6] Otsuka, T., Nakadai, K., Takahashi, T., Komatani, K., Ogata, T. and Okuno, H. G.: Design and implementation of two-level synchronization for interactive music robot, *Proc. AAAI*, (2010).
[7] Cont, A.: A coupled duration-focused architecture for real-time music-to-score alignment, *IEEE, Trans. on Pattern Analysis and Machine Intelligence*, vol.32, no.6, pp974–987, (2010).
[8] Yu, S. Z.: Hidden semi-Markov models, ” *Artificial Intelligence*, vol.174, pp.215–243, (2010).
[9] Joder, C., Essid, S., Richard, G.: A comparative study of tonal acoustic features for a symbolic level music to score alignment, *Proc. ICASSP*, pp.409–412, (2010).
[10] Welch, G. and Bishop, G.: An introduction to the Kalman filter, (2006).
[11] Emiya, V., Badeau, R. and David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle, *IEEE Trans. on Audio, Speech, and Language Processing*, vol.18, no.6, pp.1643–1654, (2010).
[12] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC music database: Popular, classical, and jazz music database, *Proc. ISMIR*, pp.287–288 (2002).

ⓒ 2012 Information Processing Society of Japan