楽曲パート混合オーディオ同士の楽譜なしアライメント手法

前澤 陽[1,a)]　奥乃 博[2,b)]

概要：本稿では，同一の楽曲を演奏する複数の音楽音響信号のうち，それぞれが楽譜の一部のパートのみを演奏した場合における，オーディオアライメント手法 – 楽曲パート混合オーディオアライメント – について報告する．本手法では，音楽音響信号を 3 つの階層 – (1) 対象となる音響信号群を構成するスペクトルテンプレートの集合，(2) スペクトルテンプレート組合せの時系列，(3) 各音響信号で出現する (2) の構成要素– で表現する．具体的には，時系列は Left-to-right 隠れマルコフモデル (LRHMM) を用い，集合の部分集合の概念を階層ディリクレ過程で表現し，スペクトルテンプレートを多項分布として表現する．評価実験から，提案手法は，演奏されるパートの違いに対して，ロバストであることが示された．

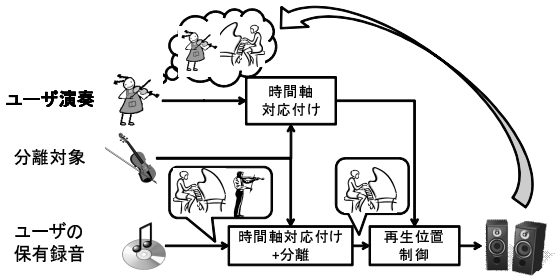

図 1　本研究のブロック図．分離対象とする音源に対し，ユーザが入力する音響信号と，ユーザの保有音源の位置対応を求める．分離対象音源の情報を元に，保有録音から所望のパートを消したオーディオを生成する．これを再生することによって，保有録音と共演している気分を楽しめる．

1. はじめに

「一緒に弾く」ことは，楽器演奏の大きな楽しみである．例えば，バイオリンとピアノのための楽曲は，バイオリン単体やピアノ単体ではなく，アンサンブルとして弾いたほうが楽しいだろう．また，伴奏する相手が，ユーザが憧れとする名奏者であれば，楽器演奏の楽しみは，より増すであろう．本研究は，ユーザが，憧れの演奏者と「一緒に弾いている」気分を楽しめるようにすることを目的とする．

このような目的を達成するために，**図 1** に示すようなシステムを考える．入力として，ユーザの実時間演奏録音 (「ユーザ演奏」)，ユーザが弾こうとしているパートの情報 (「分離対象」)，ユーザが保有している演奏録音 (「保有録音」) の三つを与える．すると，本手法では，保有録音から分離対象を除去したオーディオを用意し，このオーディオを，ユーザ演奏に同期して再生する．分離対象が除去された保有録音とユーザ演奏を同期させるには，ユーザ演奏と分離対象の同期情報と，分離対象から保有録音の同期情報を組み合わせればよい．これにより，ユーザの演奏に同期された伴奏音が再生できる．

このような実施形態において重要なのは，何を「分離対象」とするかである．分離対象の，入手の難易や汎用性によって，システムの利便性や汎用性が決まるためである．例えば，分離対象を電子楽譜表現として持つことが出来れば，楽譜表現に基づく，オーディオとの時系列対応付け [1–5] や，音源分離 [6–9] が可能となる．しかし，このような方法で楽しめるコンテンツは，電子楽譜が入手できる楽曲に限定される．また，分離対象を用意する方法として，特定の楽器に着目することも考えられる [10]．しかし，クラシック楽曲では，同じ楽器が複数のパートを演奏することは多々ある．そのため，本来残しておきたいパートまで消されてしまう恐れがある．例えば，大多数の弦楽四重奏では，二本のヴァイオリンパートが使用されているため，バイオリンパートを分離対象にした場合，消すべきパートが曖昧になる．

そこで，我々は分離対象として，ユーザが演奏した，楽曲全体の音響信号を用いることを考える．このような分離対象の設定方法は，三点のメリットがある．一点目に，ユーザが演奏できる保有録音はすべて扱える．二点目に，ユーザへの負担が少ない．ユーザは，練習時に全体を通して弾く際の音を，録音するだけでよいためである．三点目に，分離対象が明確である．

このような形で分離対象を指定する場合に問題になるのが，保有録音と分離対象音のアライメントである．ユーザ演奏と分離対象のアライメントは，従来法 [11] を用いればよい．また，保有録音と分離対象のアライメントが取れたとするならば，保有録音の音源分離には，従来法 [12] を用いればよい．しかし，保有録音と分離対象のアライメントは，従来のオーディオアライメントで算出できない．なぜならば，従来のオーディオアライメントは，二つの音響信号が近いことを想定しているが，単一パートの楽曲音で

[1] ヤマハ株式会社
Yamaha Corporation
[2] 京都大学
Kyoto University
[a)] akira_maezawa@gmx.yamaha.com
[b)] okuno@i.kyoto-u.ac.jp

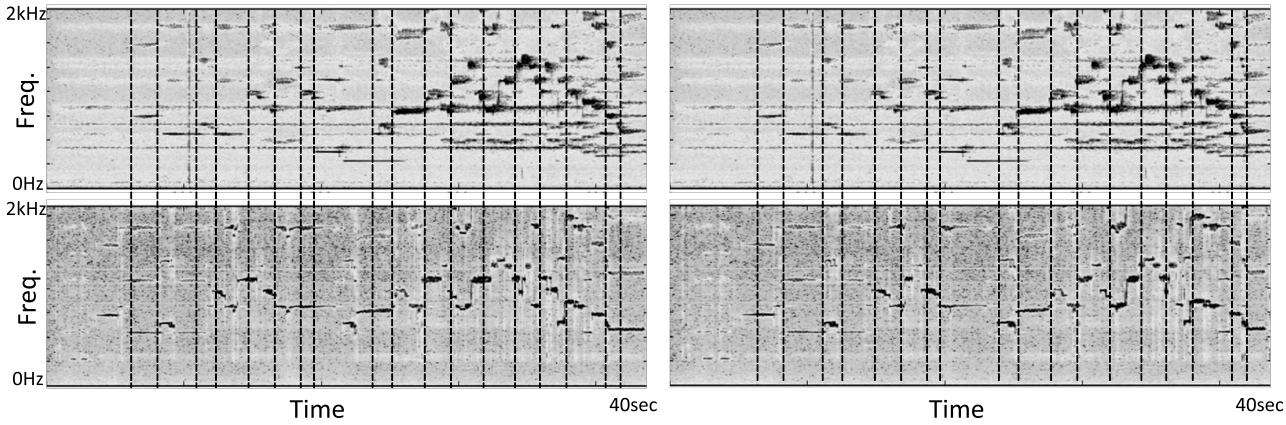

図 2 パート混合オーディオアライメントの一例．シベリウスヴァイオリン協奏曲の先頭 40 秒において，オイストラフとフィラデルフィア交響楽団の演奏（上図）に対し，筆者が演奏したソロパート（下図）をアラインさせる（左：アライン前，右：アライン後）．破線は，オイストラフの録音上でいくつかのオンセットを選び，重ねて表示している．

ある分離対象と，複数パートの混合音である保有録音は，音響信号としての乖離が激しいためである．分離対象と保有録音のアライメントが取れないため，従来法 [12] では，ユーザが，保有録音に同期して演奏をする必要があった．

そこで，本稿では，「保有録音と分離対象音のアライメント」の上位概念である，「楽曲パート混合のオーディオアライメント」という新しい音楽音響信号同士のアライメント手法を提案する．「楽曲パート混合のオーディオアライメント」とは，二つ以上の音響信号が，同一の楽譜表現のうち，それぞれ，互いに素でないような部分集合を演奏した場合のアライメントのことと定義する．例えば，ヴァイオリン，ヴィオラ，チェロから構成された楽曲に対し，ヴァイオリンとヴィオラパートが演奏された音響信号と，ヴァイオリンとチェロパートが演奏された音響信号同士を，これらに共通するヴァイオリンパートに着目してアライメントを行うようなタスクを指す．**図 2** には所望とする出力の一例を図示し，**図 3** に概念を図示する．

2. 定式化

本節では，まず，我々はパート混合アライメントの定式化における概念を直感的に説明する．次に，その概念を具体化した確率モデルを定式化し，等価である推論可能なモデルを導出する．

2.1 本手法の概念

本手法は，**図 4** に表すように，パート混合オーディオを三階層として表現する．この図では，ピアノ四重奏の楽譜に対して，3 つのパート混合オーディオ（「ドキュメント」）が存在すると仮定する．具体的には，第 1 ヴァイオリンのみが存在する信号（Doc. 1），第 1 ヴァイオリンとチェロのみが存在する信号（Doc. 2），そして，全パートが存在するオーディオ（Doc. 3）である．ここで，これらの音響信号は，それぞれ違うテンポで演奏されているが，演奏される楽譜上の位置順序は同じであると仮定する．

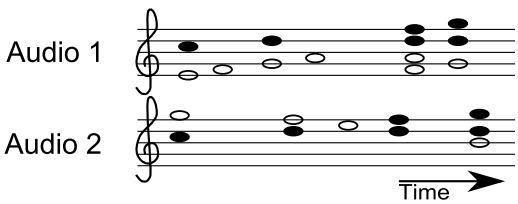

図 3 パート混合オーディオアライメントの概念．Audio 1 と 2 は，同じ楽譜の部分集合を演奏したとすると，パート混合オーディオアライメントは，両者で共通して演奏されている箇所（塗りつぶされた符頭）に着目し，時間軸対応付けをとる．

最上位の階層（「グローバルレベル」）は，全ての録音の構成要素となるようなスペクトル成分を保持している．このように，楽曲を表現するスペクトルの構成要素を「グローバル基底」と呼び，グローバル基底の組み合わせを「和音」と呼ぶ．中間層（「状態レベル」）は，励起される和音の時系列を管理する．すなわち，抽象的な楽譜表現であると言える．最下層（「ドキュメントレベル」）では，各ドキュメントの各状態において，状態レベルで定義された和音のうち，どの基底が励起されているのかを選ぶ．ここで，各階層において励起される基底の数は，特に定められていないことに注意されたい．また，各ドキュメントの各状態で励起される基底は，状態レベルで定義された和音の部分集合であり，その和音は，グローバル基底の部分集合であることに注意されたい．

このような，部分集合の性質を持つ三階層と，状態の時系列から構成されるコンセプトを，推論可能な確率モデルとして表記したい．そこで，時系列の記述に隠れマルコフモデル（HMM），部分集合の記述に階層ディリクレ過程（Hierarchical Dirichlet Process; HDP）を使用することを考える．

2.2 HDP と HMM によるモデル化

まず，確率モデルを直感的に理解するのに必要最小限な，ディリクレ過程と階層ディリクレ過程の概念を，定性的に説明する．興味のある読者は [13] 等を参照されたい．

ディリクレ過程 (DP) とは，直感的に言えば，加算無限の目を持つサイコロが与えられ，各目に対して，点が割り当てられているようなモデルである．ある目の出やすさは，その目の観測数におおよそ比例し，未観測の目を観測する

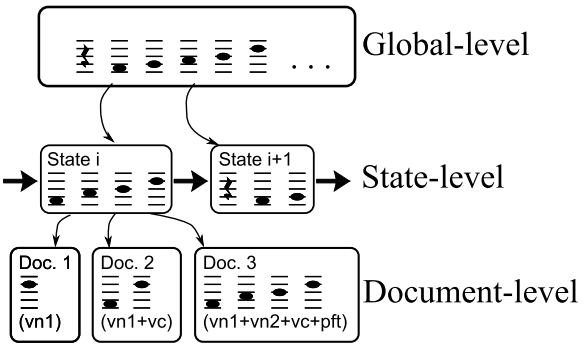

図 4 本手法の概念図．分かりやすさのため，スペクトルの基底を音符として表現している．

確率は，「集中度」というパラメータで制御される．また，各目に割り当てられる点は，「基底測度」と呼ばれる関数に従って生成される．基底測度は，別のモデルの事前分布であることが多い．本来，目の数は無限だが，すでに観測した目が，より観測されやすくなる性質がある為，実際に観測する目の数は，限られている．

次に，DP を基底測度とした，DP を考える．このような DP のことを「階層 DP (HDP)」と呼び，基底測度となる DP を「親 DP」，親 DP を基底測度とする DP を「子 DP」と呼ぶ．HDP では，子 DP が持つサイコロの目に対して，親 DP のサイコロの目を，親 DP のサイコロの重みに従って選定する．すると，選定された親 DP の目に関連付けられた点を，子 DP の目に割り当てる．つまり，子 DP が出力しうる点は，親 DP が出力できる点の部分集合となり，それぞれの点の出やすさは，子 DP におけるサイコロの重みに依存する．特に，複数の子 DP が，一つの親 DP を基底測度として与えられた場合，それぞれの子 DP は，互いに独立な，親 DP の部分集合を表現する．

これらを踏まえ，本手法では，グローバル基底を，グローバル基底の事前分布 H が基底測度となる DP G_0 として，モデル化する．つまり，サイコロの各目に対して，グローバル基底が割り当てられているように設計する．なお，H の形については後述する．このような DP は，森羅万象の振幅スペクトルを表現できる．しかし，サイコロの目が出ればその目が一層出やすくなるため，事実上，少数のグローバル基底しか観測されない．このようなモデルの事後分布を推定すると，観測確率が一定値以上であるグローバル基底の種類は，データの複雑さと集中度に応じて変化する．また，状態 s で励起される基底は，G_0 を基底測度とした HDP G_s とする．また，D 個のあるドキュメントのうち，d 番目のドキュメントにおいて，状態 s で励起される基底を，状態レベルの DP を基底測度とした HDP $G_{s,d}$ としてモデル化する．つまり，次のようなモデルを考える：

$$G_0 \sim \mathrm{DP}\,(\alpha, H) \tag{1}$$

$$G_s \sim \mathrm{DP}\,(\beta, G_0) \tag{2}$$

$$G_{s,d} \sim \mathrm{DP}\,(\gamma, G_s) \tag{3}$$

$\mathrm{DP}\,(\alpha, H)$ は，集中度 α，基底測度 H の DP を指す．

また，s の時系列を $Z\,(d,t)$ とし，状態数 S の Left-to-Right HMM (LRHMM) としてモデル化する：

$$Z(d, 1 \cdots T_d) \sim \mathrm{LRHMM}(\pi, \tau) \tag{4}$$

$\mathrm{LRHMM}(\pi,\tau)$ は，初期状態の確率が π で，状態遷移確率が τ の LRHMM を指し，$Z(d,t)$ は D 個の LRHMM から構成され，ドキュメント d の状態系列を保持する．また，全てのドキュメントは，同じ状態で終わるという制約（「強制アライメント」という）をモデル化するため，$Z^{(s)}(d, T_d) = S$ と制約する．ここで，T_d はドキュメント d の長さを示す．

2.3 信号観測モデル

$X(d,t,f)$ を，時間周波数ビン t, f（$t \le T_d$，$f \le F$）で評価したドキュメント d のパワースペクトログラムとし，これを，ドキュメント d のビン t, f の観測数と見なす．すると，各ドキュメント d の時刻 t におけるスペクトルは，多項分布 $\phi_f(d,t)$ に従い生成されると考えられる．この $\phi_f(d,t)$ は，ドキュメント d が時刻 t で用いるドキュメントレベルの DP $G_{s,d}$ に依存する．また，$G_{s,d}$ の状態 s は，状態の時系列 $Z(d,t)$ に依存する．つまり，次のようなモデルを考える：

$$\phi_f(d,t) \sim G_{Z(d,t),d} \tag{5}$$

ϕ は多項分布であるため，$G_{s,d}$ が参照する G_0 の基底測度 H は，多項分布の事前分布であることが望ましい．そこで，H を F 次元のディリクレ分布 $\mathrm{Dir}\,(g_{f,0})$ とする．ディリクレ分布以外に，調波構造といった明示的な制約を持った事前分布 [14] も考えられる．

ここで，$X(d,f,t)$ の観測回数を管理するため，変数 $C(d,c,f,t)$ を導入する．この変数は，ドキュメント d の時間周波数ビン f,t を離散化した時，c 番目の観測値が存在する，ということを示し，$c \le X(d,f,t)$ の場合のみ定義され，その値は 1 である．すなわち $\sum_c C(d,c,f,t) = X(d,t,f)$ である．これらを踏まえ，次のような観測モデルを考える：

$$C(d,c,f,t) \sim \mathrm{Mult}(\phi_f(d,t)) \tag{6}$$

ここで，$\mathrm{Mult}(\cdot)$ とは多項分布のことを指す．**図 5** に，本手法のグラフィカルモデルを図示する．

2.4 Sethuraman の棒折過程による共役モデルの構築

このようなモデルの事後分布を求めるために，変分ベイズ法を適用したい [15]．変分ベイズは，共役系（事前分布が尤度の共役事前分布であること）であると，推論が簡単である．しかし，このままでは，本手法は共役ではない．特に，HDP や $\phi_f(d,t)$ は共役形ではないので，共役な形に書きかえることが必要である．そこで，本モデルと等価な，共役なモデルを定式化する．具体的には，HDP を Sethuraman の棒折過程 [16] を用いた方法に置き換えることを考える．

まず，基底測度から，$I \to \infty$ 個のグローバル基底を生成する為に，$g_f(i) \sim \mathrm{Dir}\,(g_{f,0})$ と $w^{(g)} \sim \mathrm{SBP}\,(\alpha)$ のような確率変数を生成する．$\mathrm{SBP}\,(\alpha)$ とは棒折過程のことを指し，$w_i^{(g)}$ を，まず $\xi_i^{(g)} \sim \mathrm{Beta}\,(1,\alpha)$ と生成し，次に $w_i^{(g)} = \xi_i^{(g)} \prod_{i'}^{i-1}(1-\xi_{i'}^{(g)})$ とすることにより，生成される確率変数のことを指す．$w_i^{(g)}$ は，長さ 1 の棒を分割し，再帰的に，分割された片方の棒を二分割することによって得られると見なせるため，棒折過程と呼ばれる．

次に，各状態 s において，g を基底測度とするディリ

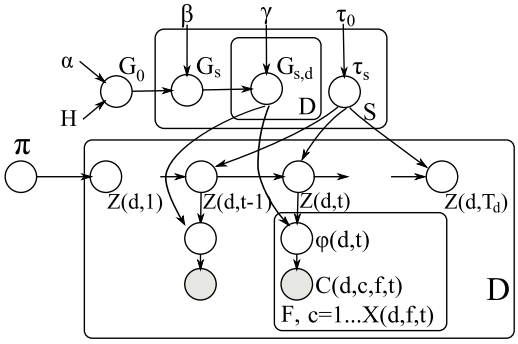

図 5 提案手法のグラフィカルモデル.

クレ過程から $J \to \infty$ 個のインジケータ変数 $Z^{(A)}(s,j)$ を生成する．これは，状態 s における和音の j 番目の構成要素が，グローバル基底のうち，どのインデックスのものを指すかを表している変数である．つまり，G_s におけるサイコロの，j 番目の目が，グローバル基底における，何番目のサイコロの目に割り当てられているかを示す．このような変数は，まず $Z^{(A)}(s,j) \sim \mathrm{Mult}(w^{(g)})$ とし，$w^{(A)}(s) \sim \mathrm{SBP}(\beta)$ とすることにより生成できる．$\mathrm{SBP}(\beta)$ は先程と同様に，$\xi_j^{(A)}(s) \sim \mathrm{Beta}(1,\beta)$ とし，$w_j^{(A)}(s) = \xi_j^{(A)}(s)\prod_{j'}^{j-1}(1-\xi_{j'}^{(A)}(s))$ とすることにより生成できる．

次に，各ドキュメント d の状態 s において，前述した状態レベルで定義された DP を基底測度とする DP を定義する．そこで，$K \to \infty$ 個のインジケータ変数 $Z^{(L)}(d,s,k)$ を生成する．これは，ドキュメント d における状態 s が生成する k 番目の基底が，状態 s における和音の，どの構成音を指すかを示す変数である．つまり，$G_{s,d}$ における，k 番目のサイコロの目が，G_s における，何番目のサイコロの目に割り当てられているかを示す変数である．具体的には，$Z^{(L)}(d,s,k) \sim \mathrm{Mult}(w^{(A)}(s))$ とし $w^{(L)}(d,s) \sim \mathrm{SBP}(\gamma)$ とする．つまり $\xi_k^{(L)}(d,s) \sim \mathrm{Beta}(1,\gamma)$ とし $w_k^{(L)}(d,s) = \xi_k^{(L)}(d,s)\prod_{k'}^{k-1}(1-\xi_{k'}^{(L)}(d,s))$ とする．

次に，スペクトルの観測 $C(d,c,f,t)$ が，$Z^{(S)}(d,t)$ が与えられた時に K 個のドキュメント単位の基底のうちの一つから生起されたと考える．そこで，$C(d,c,f,t)$ が，$G_{s,d}$ の，どのサイコロの目から生成されたかを表す潜在変数 $Z^{(X)}$ を導入する：

$$Z^{(X)}(d,c,f,t) \sim \mathrm{Mult}\Big(w^{(L)}(d,Z^{(S)}(d,t))\Big) \qquad (7)$$

この潜在変数を元に，観測尤度を次のようにモデル化する：

$$C(d,c,f,t) \sim \mathrm{Mult}\Big[g\Big(Z^{(A)}\Big(s,Z^{(L)}\Big(d,s,Z^{(X)}(d,c,f,t)\Big)\Big)\Big)\Big] \qquad (8)$$

ただし，$s = Z^{(S)}(d,t)$ とした．この式は，式 6 を，サイコロの目を割り当てる変数 $Z^{(L,X,A)}$ を通じて，グローバル基底が間接参照される形に，置き換えたものであることが分かる．

最後に，各 d における状態系列 $Z^{(S)}(d,t)$ を LRHMM としてモデル化する．具体的には，$Z^{(S)}(d,1) \sim \pi_0$ とし，$Z^{(S)}(d,t) \sim \tau(Z^{(S)}(d,t-1))$ とする．ここで，π_0 と $\tau(s)$ はそれぞれ初期状態確率と状態 s における状態遷移確率である．LRHMM であるため，π_0 は最初のインデックス以外が 0 である超パラメータを持つ Dirichlet 分布から生成される．また，τ_s は，s と $s+1$ 番目以外の要素が 0 である超パラメータを持つ Dirichlet 分布から生成される．

表 1 実験で用いられた設定.

変数名	値
α,β,γ	100, 50, 50
S,I,J,K	$\min(T_1,T_2)$, 95, 20, 10
$w_{f,0}(i)$	$100e^{-\left(f-440\times 2^{\frac{i-69}{12}}\right)^2}$
π_0	最初のインデックスは 1，それ以外は 0
$\tau_0(s)$	インデックス s と $s+1$ が 1，それ以外は 0

ここで，今までに導入した潜在変数・インジケータを 1-of-K の二値変数とする．つまり，例えば $Z^{(S)}(d,t) = s'$ は，$Z_{s'}^{(S)}(d,t) = 1$ で，それ以外の要素が 0 であるような変数として表記する．すると，完全対数尤度は，定数項を無視すると次のように与えられる：

$$\begin{aligned}
&\sum_{d,t,f,c,s,i,j,k} Z_i^{(A)}(s,j)Z_j^{(L)}(d,s,k)Z_k^{(X)}(d,c,f,s)Z_s^{(S)}(d,t)\log g_f(i) \\
&\quad + \sum_{s,i,j} Z_i^{(A)}(s,j)\log w_i^{(g)} + \sum_{d,s,j,k} Z_j^{(L)}(d,s,k)\log w_j^{(A)}(s) \\
&\quad + \sum_{d,t,f,c,s,k} Z_k^{(X)}(d,c,f,s)Z_s^{(S)}(d,t)\log w_k^{(L)}(d,s) \\
&+ \sum_{d,t,s,s'} Z_s^{(S)}(d,t-1)Z_{s'}^{(S)}(d,t)\log\tau_{s,s'} + \sum_{d,t,s} Z_s^{(S)}(d,0)\log\pi_s \\
&\quad + \log \mathrm{SBP}\left(w_i^{(g)}|\alpha\right) + \sum_s \log \mathrm{SBP}\left(w_j^{(A)}(s)|\beta\right) \\
&\quad + \sum_{d,s} \log \mathrm{SBP}\left(w_k^{(L)}(d,s)|\gamma\right) + \log \mathrm{Dir}\left(\pi|\pi_0\right) \\
&\quad + \sum_s \log \mathrm{Dir}\left(\tau_s|\tau_0\right) + \sum_i \log \mathrm{Dir}\left(g_f(i)|g_{f,0}\right) \qquad (9)
\end{aligned}$$

強制アライメントを行うため，$Z_S^{(S)}(d,T_d) = 1$ とし，$\hat{s} < S$ となるような $\hat{s}$ においては $Z_{\hat{s}}^{(S)}(d,T_d) = 0$ とする．

このモデルは共役であるので，変分ベイズ法により事後分布の近似を効率的に求めることが出来る．特に，$Z^{(S)}$ については，前向き後ろ向きアルゴリズムを組み合わせることにより効率的な推論が可能になる．I, J, K は無限であるが，その有限近似 [16] を用いることで，有限な計算リソースでも推定が可能になる．紙面の制約上，導出は割愛する．興味のある読者は，[15,16] 等を参考にされたい．

事後分布が求まったら，状態系列の最大事後確率 (MAP) 推定値 $\hat{s}(d,t) = \arg\max_s \langle Z_s^{(S)}(d,t)\rangle$ を，全てのドキュメントに対して求める．すると，任意の状態を取る時刻を，全てのドキュメントに対して求めることにより，アライメントが求まる．

3. 実験と考察

評価のため，提案手法と，従来の DTW に基づくオーディオアライメント手法 [11,17] の亜種（以下「cos-DTW」）を比較する．以下，「完全オーディオ」を，楽譜に記載されている全てのパートが含まれたオーディオとし，「ソロオーディオ」を，単一パートのみを演奏したオーディオとする．まず，両手法の通常のオーディオアライメント（完全オーディオ対完全オーディオのアライメント）における性能を

表 2 ベースライン（「cos-DTW」）と提案手法（「Proposed」）におけるアライメント誤差の比較．「full-to-full」は完全オーディオ対完全オーディオ，「solo-to-full」はメロディーパート対完全オーディオ，「parts-to-parts」は，パート混合対パート混合における，あらゆるパートの組み合わせで得られた結果の平均，「parts-to-full」は，パート混合対完全パートにおける，あらゆるパートの組み合わせで得られた結果の平均である．

楽曲	パート構成	手法	誤差 <1.0s	誤差 <2.0s	誤差 <5.0s	誤差 <10.0s
J.S. Bach, BWV847 フーガ全体	Piano LH + Piano RH	cos-DTW(full-to-full)	89%	99%	100%	100%
		Proposed (full-to-full)	79%	97%	99%	100%
		cos-DTW (RH-to-full)	69%	83%	87%	88%
		Proposed (RH-to-full)	74%	89%	94%	94%
F. Chopin, Op. 22 Polonaise 先頭 16 小節	Piano LH+ Piano RH	cos-DTW(full-to-full)	92%	100%	100%	100%
		Proposed (full-to-full)	86%	100%	100%	100%
		cos-DTW (RH-to-full)	85%	98%	100%	100%
		Proposed (RH-to-full)	87%	99%	100%	100%
J. Brahms, Op. 40 1 楽章 先頭 32 小節	French Horn+ Violin+ Piano LH+ Piano RH	cos-DTW(full-to-full)	96%	99%	100%	100%
		Proposed (full-to-full)	81%	98%	100%	100%
		cos-DTW (parts-to-parts)	46%	54%	60%	64%
		Proposed (parts-to-parts)	51%	71%	92%	98%
		cos-DTW (parts-to-full)	50%	56%	60%	63%
		Proposed (parts-to-full)	59%	78%	92%	96%
P. Tchaikovsky, Op. 35 2 楽章 8–34 小節	Violin Solo + Orchestra	cos-DTW(full-to-full)	94%	98%	99%	100%
		Proposed (full-to-full)	77%	96%	100%	100%
		cos-DTW (solo-to-full)	52%	60%	64%	69%
		Proposed (solo-to-full)	44%	74%	98%	100%

比較する．次に，パート混合オーディオアライメントにおける性能を比較する．

まず，**表 2** に記された楽曲の SMF に対して，各パートの音響信号をソフトウエアシンセサイザーで生成した．なお，ピアノパートは，右手（RH）と左手（LH）を別パートとして生成した．次に，各楽曲の各パートに対して，再生速度を 20%遅らせた信号をタイムストレッチ技術で用意した．次に，通常の速度で再生された完全オーディオに対して，低速再生された完全オーディオと低速再生されたソロオーディオの，2 種類のアライメントを求めた．最後に，これら 2 種類のアライメントに対して，累計絶対誤差を求めた．ただし，Brahms Op. 40 では，全てのパートの組み合わせに対する，パート混合オーディオと完全オーディオをアラインさせた際の平均を求めた．また，Brahms Op. 40 では，全ての 2 パート混合と 3 パート混合同士に対して，最低一つの共通パートが演奏されているような，パート混合同士のアライメントを計算した．

$X(d,t,f)$ の算出には，サンプリング周波数 44.1kHz で生成された各音響信号に対し，Bartlett-Hanning 窓を適用し，フレーム長 8192 サンプル，ホップサイズ 4096 サンプルで振幅スペクトログラムを生成し，周波数成分 2kHz 以上の成分を破棄した．また，本手法は**表 1** に示すようなパラメータを用い，π，τ と $g_f(i)$ 以外の変数における事後分布が推定された．

cos-DTW におけるアライメントは，コサイン距離をスペクトル同士の距離としたときの，スペクトル間の累計距離を最小化するような経路を，DTW で求めた．ただし，通常のオーディオアライメントで用いられる DTW の状態遷移は，不当に低性能になりやすいため，状態遷移を LRHMM のそれと同一のものにした．また，提案手法と cos-DTW に使う特徴量は振幅スペクトルのみと統一させた．つまり，近年のオーディオアライメント手法のような高次な特徴量 [11] は，敢えて入れていない．そのため，cos-DTW は，現在のオーディオアライメント手法を，代表するものではないことに注意されたい．これらを踏まえると，cos-DTW は，HMM の Viterbi アルゴリズムと見なすことができる．HMM の状態は，再生時間が長い音響信号の，各時刻一つ一つで定義され，その観測尤度は，各時刻におけるパワースペクトルを位置パラメータに持ち，集中度 1 の von Mises-Fisher 分布に従うものと見なすことができる．つまり，本実験での性能差は，時系列モデルの良し悪しには強く起因せず，スペクトルのモデル化の違いに起因することになる．

実験結果を表 2 に示す．この結果から，完全オーディオアライメントでは，提案手法は，比較手法と比べ同程度か，多少低い性能であることが分かる．この理由として，(1) DTW は大局的に最適であるが，提案手法は局所解に陥ることがある，(2) コサイン距離は，同一音源に対するスペクトルの距離尺度として，よい尺度である，という二点が考えられる．一方，パート混合オーディオと完全オーディオをアラインする際には，提案手法の方が比較手法よりもエラーが少なくなる．とりわけ，致命的となるような 2 秒以上のエラーが緩和されている．コサイン距離の場合，メロディーパート単体のスペクトルと，完全オーディオのスペクトル上の乖離が激しいため，アライメントが失敗する．一方，提案手法は，二つの信号で共通する箇所に着目するため，性能低下を抑えることができると考えられる．この点は，**図 6** によく現れている．バイオリンソロとバイオリン＋オーケストラのように，スペクトル上の違いが多すぎる場合，cos-DTW の追従は失敗する．一方で，提案手法

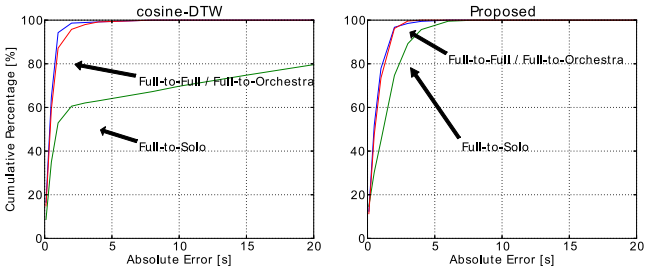

図 6 Tchaikovsky Op. 35 のソロパートに対する完全オーディオ/オーケストラパートに対する完全オーディオのアライメントを，cos-DTW（左）と提案手法（右）で算出した結果.

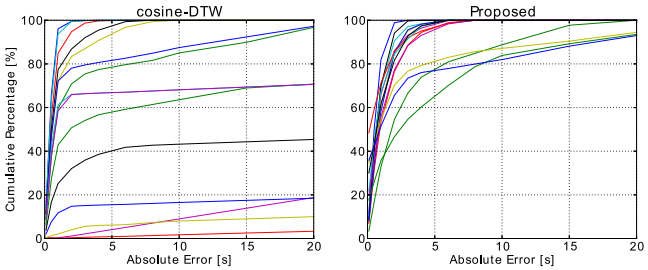

図 7 Brahms Op. 40 の，2 パート/3 パート混合同士のアライメントを，cos-DTW（左）と提案手法（右）で算出した結果.

は楽曲を通して追従ができているため，提案手法のような三階層のモデル化の有効性が示唆される．また，Brahms Op.40 において，パート混合オーディオ間のアライメントを行った結果を**図 7** に示す．この図からも，提案手法は，パート構成の違いに対してロバストなアライメントが実現できていることが分かる．

ただし，提案手法は，そもそものアライメント精度向上の余地があるだろう．特に，状態系列が，必要以上に一つの状態に長く留まる場合に，ミスが起こりやすいことが予備実験で確認された．よって，セミマルコフ的な状態遷移 [2] を導入することにより，精度向上が期待できる．

4. まとめ

本稿では，パート混合オーディオアライメント手法を提案した．パート混合オーディオアライメントとは，複数の音響信号が，同一楽譜の，互いに疎ではない一部分を演奏した場合における，時間的な対応付けを求めるタスクとした．アライメントを行うため，パート混合オーディオの集まりの生成モデルを考え，三階層ディリクレ過程から生起されるヒストグラムの時系列としてモデル化し，その事後分布を求めた．評価の結果，演奏されるパートの種類の違いに対してロバストなアライメントが実現できていることが確認された．

今後の課題として，音源分離手法への応用，局所解に陥りにくい推論方法の確立，信号の生成モデルという観点で見た時により妥当なモデルの確立，アタックのような高次情報の導入，セミマルコフ的な時系列モデルの導入などがある．また，オンラインで本モデルを解ければ，分離対象とユーザ演奏を同一のものとして扱えるため，本手法のオンライン化も重要な課題である．

また，他の問題に，本手法の枠組みを適用することも考えられる．例えば，楽譜追従タスクにおいて，カデンツァや演奏ミスを，楽譜表現からの逸脱として見なすのではなく，カデンツァや演奏ミスを含む「楽譜」から，片方（SMF）はミスやカデンツァが欠落され，もう片方（演奏）にはミスやカデンツァを含む楽譜が生成されると考えると，楽譜追従における「楽譜と音響信号のミスマッチ」に新たな解決案を提案できるだろう．

参考文献

[1] 前澤 陽，糸山克寿，尾形哲也，奥乃 博：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，情報処理学会音楽情報科学研究会 [2011-MUS-91] (2011).
[2] Cont, A.: A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment, *IEEE PAMI*, Vol. 32, No. 6, pp. 974–987 (2010).
[3] Dannenberg, R. B. and Raphael, C.: Music score alignment and computer accompaniment, *CACM*, Vol. 49, No. 8, pp. 38–43 (2006).
[4] Müller, M. and Ewert, S.: Towards Timbre-Invariant Audio Features for Harmony-Based Music, *IEEE TASLP*, Vol. 18, No. 3, pp. 649–662 (2010).
[5] Joder, C., Essid, S. and Richard, G.: A conditional random field viewpoint of symbolic audio-to-score matching, *ACMM*, pp. 871–874 (2010).
[6] Ewert, S. and Müller, M.: Using score-informed constraints for NMF-based source separation, *ICASSP*, pp. 129–132 (2012).
[7] Han, Y. and Raphael, C.: Informed Source Separation of Orchestra and Soloist, *ISMIR*, pp. 315–320 (2010).
[8] Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions, 情報処理学会論文誌，Vol. 50, No. 7, pp. 1757–1767 (2009).
[9] Hennequin, R., David, B. and Badeau, R.: Score informed audio source separation using a parametric model of non-negative spectrogram, *ICASSP*, pp. 45–48 (2011).
[10] 藤原弘将，後藤真孝：混合音中の歌声スペクトル包絡推定に基づく歌声の声質変換手法，情報処理学会音楽情報科学研究会 [2010-MUS-86] (2010).
[11] Dixon, S.: MATCH: A music alignment tool chest, *ISMIR*, pp. 492–497 (2005).
[12] Smaragdis, P. and Mysore, G.: Separation by humming: User-guided sound extraction from monophonic mixtures, *WASPAA*, pp. 69–72 (2009).
[13] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M.: Hierarchical Dirichlet Processes, *JASA*, Vol. 101, No. 476, pp. 1566–1581 (2006).
[14] Yoshii, K. and Goto, M.: A Nonparametric Bayesian Multipitch Analyzer Based on Infinite Latent Harmonic Allocation, *IEEE TASLP*, Vol. 20, No. 3, pp. 717–730 (2012).
[15] Beal, M. J.: Variational Algorithms for Approximate Bayesian Inference, PhD Thesis, University College London (2003).
[16] Wang, C., Paisley, J. W. and Blei, D. M.: Online Variational Inference for the Hierarchical Dirichlet Process, *JMLR*, Vol. 15, pp. 752–760 (2011).
[17] Dannenberg, R. B. and Hu, N.: Polyphonic Audio Matching for Score Following and Intelligent Audio Editors, *ICMC* (2003).