情報処理学会研究報告
IPSJ SIG Technical Report

無限状態スペクトルモデルによる音楽音響信号の解析

中野 允裕[†1]　ルルー ジョナトン[†2]
亀岡 弘和[†2]
小野 順貴[†1]　嵯峨山 茂樹[†1]

本報告では, 多重音から時間変化する楽器音スペクトルを学習することを目的とした音楽音響信号解析のための無限状態スペクトルモデルを提案する. 歌声や弦楽器におけるヴィブラートなどの時間変化するスペクトルをディリクレ過程によって生成されたモデルとして音楽信号を表現する. 提案手法は非負値行列分解 (NMF) の拡張であり, 本報告では 事後確率最大化の観点から効率的なアルゴリズムを導出し, 実際に音楽信号に適用した実験例を示す.

Music Signal Analysis with Infinite-State Spectrum Model

Masahiro Nakano,[†1] Jonathan Le Roux ,[†2]
Hirokazu Kameoka,[†2] Nobutaka Ono[†1]
and Shigeki Sagayama[†1]

This paper presents infinite-state spectrum model to learn time-varying spectrum in polyphonic music signals. Time-varying spectrum , such as vibrato of singing voices or the stringed instruments, is represented by infinite-state model by the Dirichlet process. We describe our extension of nonnegative matrix factorization (NMF) in a statistical framework. An efficient algorithm for Maximum a posteriori (MAP) estimation is tested on real audio data.

†1 東京大学情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
†2 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation

1. はじめに

近年, インターネットの発達やポータブルプレーヤの普及に伴って, 一般ユーザが大量の音楽データを保有するようになり, 検索技術などそれらを管理するための技術の需要が高まってきている. また, 単に受動的に音楽を鑑賞するだけでなく, 特定の楽器の音量だけを操作したり, 曲の一部分を補正して聴きたいといった能動的音楽鑑賞の需要も生まれてきた.

これらを実現するための基礎技術として, 多重音を解析するための手法が盛んに研究されており, 近年, 特にスパース表現と呼ばれる考え方に基づく手法が注目されている. 特に信号を時間周波数表現したスペクトログラムに対して用いる非負値行列分解[1),2)] は, 多重音からの自動採譜[3)], 音源分離[4)], 音楽加工[5)] といった応用によく利用されている.

スペクトログラムに対して行う非負値行列分解は, 限られた数の基底スペクトルパターンが音量だけ変化し, それらの重ね合わせによって音楽信号のスペクトログラムが観測されるというモデルに基づいて多重音を表現していることに相当する. この時, 分解されるスペクトルパターン一つ一つが自然と楽器音一音一音に対応してくれることを期待している. しかし, ピアノにおけるアタック, ディケイ, サステイン, リリースや, 歌声, 弦楽器におけるヴィブラートのように, 実際の楽器音は時間に伴って変化するスペクトルを持っており, これらを一つの基底スペクトルパターンとして表現するモデルには限界があった.

そこで近年, 時間変化する楽器音スペクトルを表現するためのモデルの改良が行われてきた.[6)] ではソースフィルタモデルに基づいて, ARMA モデルによる時間周波数領域でのアクティベーションを持つモデルへと NMF の拡張が行われている. しかし, ピアノの打鍵時に生じるアタック音（広帯域な周波数にエネルギーを持つ）とサステイン部（調波構造を持つ）のように元々の起源の異なる時間変化にソースフィルタモデルが適したモデルかどうかははっきりしていない. これに対し, 基底スペクトルパターンの時間変化を導入したモデルも提案されている.[9)] では, Factorial hidden Markov model に基づいて時間変化するスペクトルを扱うモデルが提案されており, これは板倉齊藤距離規準の NMF の拡張と見なすことも出来る. また,[7),8)] では, 最適化問題としての β-divergence 規準の NMF に基底の状態遷移を導入したモデルが提案されている. しかし, これらの手法において, 楽器音一音高をいくつの基底で表現するかは事前に決めておく必要があった.

本報告では, 時間変化するスペクトルをスペクトルパターンの状態遷移によって表現し, かつ各楽器音を表現するための状態数を同時に推定する枠組みを提案する.

2. 無限状態スペクトルモデル

2.1 従来の有限状態スペクトルモデル

非負値行列分解 (NMF) を音楽信号に適用する場合, 一般的には振幅スペクトログラムまたはパワースペクトログラム $\mathbf{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$（ただし $\omega = 1,\cdots,\Omega$ は周波数のインデックス, $t = 1,\cdots,T$ は時間のインデックス）を基底 $\mathbf{H} = (H_{\omega,d})_{\Omega\times D} \in \mathbb{R}^{\geq 0,\Omega\times D}$ と アクティベーション $\mathbf{U} = (U_{d,t})_{D\times T} \in \mathbb{R}^{\geq 0,D\times T}$ に分解する形で用いられることが多い. すなわち

$$Y_{\omega,t} \approx \sum_d H_{\omega,d} U_{d,t} \ , \tag{1}$$

と表される. ここで D は基底ベクトル $\mathbf{h}_d = [H_{1,d},\cdots,H_{\Omega,d}]$ の数を表している. 音楽信号の場合, 基底ベクトル $\mathbf{h}_d$ は頻出のスペクトルパターンを表し, アクティベーション $\mathbf{u}_d$ は直感的にはそれぞれのスペクトルパターンに対応した音量に相当するものだと考えることが出来る. つまり, 非負値行列分解による音楽信号の分解は, 音量のみが時間変化する限られた数のスペクトルパターンの重ね合わせで表現されるモデルによって, 観測スペクトログラムを表そうとしたものであると考えることが出来る.

理想的には, 一つの基底スペクトルパターンとアクティベーションのペアで楽器の一音高を表していることを期待している. しかし, 実際の楽器音のスペクトルは非定常なため, 楽器一音高が複数の基底によって表現されてしまう場合があり, それら複数の基底を一つの音高として扱うには何らかの後処理が必要となるという問題があった. この問題に対して, 基底スペクトルパターンは時間に伴って変化するようにモデルを拡張することが提案されている[7)–9)]. 特に[7),8)] においては, 各基底スペクトルパターンは時刻 t に, ある一つの状態 $Z_{d,t} \in \mathbb{N}$ をとると見なし,

$$Y_{\omega,t} \approx \sum_d H_{\omega,d}^{(Z_{d,t})} U_{d,t} \ , \tag{2}$$

のようにモデルを拡張することが考えられている. このとき, モデルパラメータの推定はモデルと観測スペクトログラムを何らかの距離尺度を最小化するという最適化問題というアプローチ[7),8)] と, 統計的な観点からモデルパラメータを推定するアプローチ[9)] の 2 つの戦略が考えられてきた.

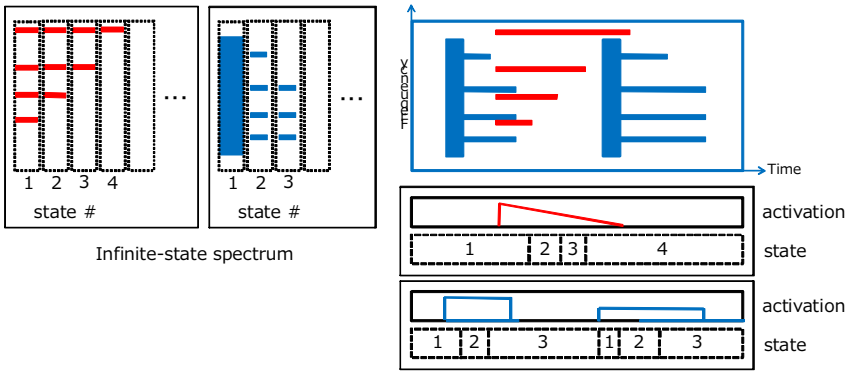


図 1 無限状態スペクトルモデルの概念図

本報告ではこれ以降, 統計的な観点からモデルを表現することにする. 観測スペクトログラムが D 個の要素 $\mathbf{C} = (C_{\omega,t,d})_{\Omega\times T\times D}$ の重ね合わせとして表現されていると考え,

$$Y_{\omega,t} = \sum_d C_{\omega,t,d} \ , \quad C_{\omega,t,d} \sim \mathrm{Poisson}(C_{\omega,t,d} \mid H_{\omega,d}^{(Z_{d,t})} U_{d,t}) \ . \tag{3}$$

というモデルを考える. これはポアソン分布の再生性から,[7),8)] において, モデルと観測の間の距離尺度を Kullback-Leibler (KL) divergence とした場合と等価である. すなわち,

$$Y_{\omega,t} \sim \mathrm{Poisson}(Y_{\omega,t} \mid \sum_d H_{\omega,d}^{(Z_{d,t})} U_{d,t}) \ . \tag{4}$$

が成り立つ. KL divergence は音源分離の際に用いられることが多い規準である[4)]. 二乗誤差規準と板倉齊藤距離規準の場合においては, 式 (3) をガウス分布, 複素ガウス分布とすることで, これ以降と同様の議論をすることが出来る[15)].

このように時間変化する基底スペクトルを導入する際, 考えなければならない問題として, 基底のとりうる状態数の決定方法がある. 従来は, 事前に全ての基底に同じ数の状態を与えていたり[7),8)], 対象に応じて（例えば, 話し声と音楽）あらかじめ基底ごとに異なる状態数を与えていた[9)]. しかし, 一般的には観測される音楽信号の事前知識がないことも多く, 必要な状態数はモデルが自動的に推定してくれることが理想的である. そこで, 本報告では, 式 (2) のモデルの状態数を自動的に推定する枠組みへと拡張する.

2.2 無限状態スペクトル

ピアノのような打鍵楽器はアタック, ディケイ, サステイン, リリースと呼ばれるように, スペクトルは 3, 4 個の状態によって表現出来ると考えられる. 一方, 歌声や弦楽器のヴィブラートのスペクトルは多様に変化するものであり, どの程度の状態数を用意しておくべきかを事前に決定することは現実的ではない. そこで, 観測データに基づき, 状態数が必要に応じて増えていく柔軟なモデルが必要である. 本報告では, 楽器音のスペクトルパターンがディリクレ過程によって生成されたと考えることによって, 従来の有限状態スペクトルモデルを状態数が自動的に推定されるような柔軟なモデルへ拡張する.

d 番目の楽器音の基底スペクトルの状態時系列 $\{Z_{d,1}, \cdots, Z_{d,T}\}$ はそれぞれ離散的な値 $1, \cdots, K$（すなわち状態のインデックス）をとる. このとき, これらの状態時系列の同時分布を $\boldsymbol{\pi}_d = \{\pi_{d,1}, \cdots, \pi_{d,K}\}$ $(\forall d, k, \pi_{d,k} \geq 0, \sum_k \pi_{d,k} = 1)$ を用いて,

$$p(Z_{d,1}, \cdots, Z_{d,T} \mid \boldsymbol{\pi}_d) = \prod_{k=1}^{K} \pi_{d,k}^{n_d^{(k)}} \quad , \quad \text{with} \quad n_d^{(k)} = \sum_{t'=1}^{T} \delta(Z_{d,t'} - k) \tag{5}$$

とする. ただし, $\delta(x-k) = 1$ iff $x = k$, otherwise 0 とし, $n_d^{(k)}$ は $Z_{d,t'} = k$ $(t' = 1, \cdots, T)$ を満たす t' の個数を表している. ここで, $\boldsymbol{\pi}_d = \{\pi_{d,1}, \cdots, \pi_{d,K}\}$ に対して, 次のような事前分布を考える.

$$p(\pi_d \mid \gamma_d) \sim \text{Dirichlet}(\pi_d \mid \gamma_d/K, \cdots, \gamma_d/K) = \frac{\Gamma(\gamma_d)}{\Gamma(\gamma_d/K)^K} \prod_{k=1}^{K} \pi_{d,k}^{\gamma_d/K-1} \quad , \tag{6}$$

ただし, $\boldsymbol{\gamma} = \{\gamma_1, \cdots, \gamma_D\}$ は正のパラメータとする. このように事前分布を定めると, $\boldsymbol{\gamma}$ に対する状態時系列の同時分布は,

$$\begin{aligned} p(Z_{d,1}, \cdots, Z_{d,T} \mid \gamma_d) &= \int p(Z_{d,1}, \cdots, Z_{d,T} \mid \boldsymbol{\pi}_d) p(\boldsymbol{\pi}_d \mid \gamma_d) d\boldsymbol{\pi}_d \\ &= \frac{\Gamma(\gamma_d)}{\Gamma(\gamma_d + T)} \prod_{k=1}^{K} \frac{\Gamma(n_d^{(k)} + \gamma_d/K)}{\Gamma(\gamma_d/K)}. \end{aligned} \tag{7}$$

と書くことが出来る. したがって, $\{Z_{d,1}, \cdots, Z_{d,t-1}, Z_{d,t+1} \cdots Z_{d,T}\}$ (以降では $\mathbf{Z}_{d,-t}$ と表現する) が与えられた時の $Z_{d,t}$ の条件付き確率は次のように与えられる.

$$p(Z_{d,t} = k \mid \mathbf{Z}_{d,-t}, \gamma_d) = \frac{n_{d,-t}^{(k)} + \gamma_d/K}{T - 1 + \gamma_d} \quad . \tag{8}$$

ただし, $n_{d,-t}^{(k)}$ は $Z_{d,t'} = k$ $(Z_{d,t'} \in \mathbf{Z}_{d,-t})$ を満たす t' の個数を表しているものとする. 式 (8) から分かるように, $Z_{d,t}$ は他の時刻においても良く選ばれている状態を取りやすくなる性質がある.

さらに, 楽器音スペクトルの状態数は必要な数だけ増やすことの出来る枠組みを導入したい. ディリクレ過程においては, 有限として考えてきた状態数を $K \to \infty$ とすることができ, $Z_{d,-t}$ が与えられた時の $Z_{d,t}$ の条件付き確率は次のように表せる.

$$p(Z_{d,t} = k \mid \mathbf{Z}_{d,-t}, \gamma_d) = \begin{cases} \dfrac{n_{d,-t}^{(k)}}{T - 1 + \gamma_d} & (n_{d,-t}^{(k)} > 0) \\ \dfrac{\gamma_d}{T - 1 + \gamma_d} & (Z_{d,t}\text{が新規の状態となる場合}) \end{cases} . \tag{9}$$

上式から分かるように, 無限状態スペクトルモデルにおいて, 各時刻に用いられる状態に着目すると, 他の時刻に多く用いられている状態ほど使われやすくなる性質がある. また, 新しい状態が用いられやすくなるか否かについてはパラメータ γ_d が影響している.

3. 事後確率最大化アルゴリズム

本章では, 提案モデルのパラメータ推定を事後確率最大化の観点から考える. すなわち, $\{\mathbf{H}, \mathbf{U}\}$ を点推定する問題として考えるが, $\{H_{\omega,d}^{(1)})_{\Omega \times D}, (H_{\omega,d}^{(2)})_{\Omega \times D}, \cdots\}$ は理論上可算無限個まで増える可能性がある. しかし, 実際に提案モデルを有限の長さの音楽に適用する場合は, 各楽器音スペクトルを表現するのに必要な状態数は有限に収まる. そこで, 十分大きな状態数で打ち切りを持つ Stick-breaking 過程[10] を考えることによって効率的なパラメータ推定アルゴリズムを導出することが出来る.

ディリクレ過程の構成法として次のような Stick-breaking 表現が知られている[11].

$$V_{d,k} \sim \text{Beta}(1, \gamma) \quad , \quad \pi_{d,k}(\mathbf{V}_d) = V_{d,k} \prod_{j=1}^{k-1} (1 - V_{d,j}) \quad , \tag{10}$$

ただし, $\mathbf{V}_d = \{V_{d,1} \cdots, V_{d,T}\}$. ここで, 打ち切りのための定数を K' として, $\forall d, p(V_{d,K} = 1) = 1$ とする. すなわち, $\forall k > K', \pi_{d,k} = 0$ とする.

次に基底とアクティベーション $\mathbf{H}, \mathbf{U}$ の事前分布について考える. もっとも明快な事前分布の選び方は, 式 (3) のポアソン分布に対する共役事前分布であるガンマ分布を選ぶことである[12]. 実際, ガンマ分布の事前分布はスパースな解を導くという報告がある[12]. しかし, 音楽のスペクトログラムに対して NMF が適用される場合, よく用いられるのは, 基底スペ

クトルパターンには事前分布を設定せず, アクティベーションをなめらかにするような制約を課す方法である[4),13)]. これは, 楽器音の音量がなめらかに変化することを利用したもので,[4),13)] で報告されているように, NMF に対する制約として最もよく用いられるスパース性を導入することなく, 音楽のスペクトログラムに対して有効に機能することが知られている. そこで, 本報告でもアクティベーション $\mathbf{U}$ にのみ, 次のような制約を導入する.

$$p(\mathbf{U}) = \prod_{d} \prod_{\omega=2}^{T} p(U_{d,t} \mid U_{d,t-1}) \ . \tag{11}$$

$p(U_{d,t} \mid U_{d,t-1})$ の選び方に関しても自由があるが, ここでは[13),14)] と同様に

$$p(U_{d,t} \mid U_{d,t-1}) = \mathrm{InverseGamma}\left(U_{d,t} \mid \beta_d, (\beta_d + 1)U_{d,t-1}\right) \ . \tag{12}$$

を用いることにする. 逆ガンマ分布以外に考えられるものとしては, ガンマ分布[14)] や[16)] のようなガウス分布がある.

これ以降, Θ を $\{\boldsymbol{\gamma}, \beta_1, \cdots, \beta_D\}$ とする. このとき, 事後分布は

$$\begin{aligned}
&\log p(\mathbf{H}, \mathbf{U} \mid \mathbf{Y}, \Theta) \\
&\quad = \log \sum_{\mathbf{Z},\mathbf{C}} \int_{\mathbf{V}} p(\mathbf{H}, \mathbf{U}, \mathbf{Z}, \mathbf{C}, \mathbf{V} \mid \mathbf{Y}, \Theta) d\mathbf{V} \\
&\quad \geq \sum_{\mathbf{Z},\mathbf{C}} \int_{\mathbf{V}} q(\mathbf{Z}, \mathbf{C}, \mathbf{V}) \log \frac{p(\mathbf{H}, \mathbf{U}, \mathbf{Z}, \mathbf{C}, \mathbf{V} \mid \mathbf{Y}, \Theta)}{q(\mathbf{Z}, \mathbf{C}, \mathbf{V})} d\mathbf{V} \\
&\quad = \sum_{\mathbf{Z},\mathbf{C}} \int_{\mathbf{V}} q(\mathbf{Z}, \mathbf{C}, \mathbf{V}) \log \frac{p(\mathbf{C} \mid \mathbf{H}, \mathbf{U}, \mathbf{Z}, \Theta) p(\mathbf{U} \mid \Theta) p(\mathbf{Z} \mid \mathbf{V}) p(\mathbf{V} \mid \Theta)}{q(\mathbf{Z}, \mathbf{C}, \mathbf{V}) p(\mathbf{Y} \mid \Theta)} d\mathbf{V} \ .
\end{aligned} \tag{13}$$

等号は $q(\mathbf{Z}, \mathbf{C}, \mathbf{V}) = p(\mathbf{Z}, \mathbf{C}, \mathbf{V} \mid \mathbf{Y}, \mathbf{H}, \mathbf{U})$ の時に成立する. したがって, 事後分布は EM アルゴリズムの原理から次のような反復計算により単調増加させることが出来る.

$$\begin{aligned}
\mathrm{E-step}: &\quad q(\mathbf{Z}, \mathbf{C}, \mathbf{V}) = p(\mathbf{Z}, \mathbf{C}, \mathbf{V} \mid \mathbf{Y}, \mathbf{H}, \mathbf{U}) \ , \\
\mathrm{M-step}: &\quad \{\mathbf{H}, \mathbf{U}\} = \operatorname*{argmax}_{\mathbf{H},\mathbf{U}} E_{q(\mathbf{Z},\mathbf{C},\mathbf{V})}[\mathcal{J}(\mathbf{H}, \mathbf{U}, \mathbf{C}, \mathbf{Z})] \ ,
\end{aligned} \tag{14}$$

ただし, $\mathcal{J}(\mathbf{H}, \mathbf{U}, \mathbf{C}, \mathbf{Z}) = \log p(\mathbf{C}|\mathbf{H}, \mathbf{U}, \mathbf{Z}) + \log p(\mathbf{U} \mid \Theta) + \log p(\mathbf{Z} \mid \mathbf{V}) + \log p(\mathbf{V} \mid \Theta)$ とし, $E_q[\cdot]$ は q のもとでの $\cdot$ の期待値を表しているものとする.

しかし, 実際に E-step の $p(\mathbf{Z}, \mathbf{C}, \mathbf{V} \mid \mathbf{Y}, \mathbf{H}, \mathbf{U})$ を計算することは困難なため,

$$q(\mathbf{Z}, \mathbf{C}, \mathbf{V}) = \prod_{d} \prod_{k=1}^{K'-1} q(V_{d,k} \mid \alpha_{d,k,1}, \alpha_{d,k,2}) \prod_{d,t} q(Z_{d,t} \mid \phi_{d,k}) \prod_{\omega,t,d} q(C_{\omega,t,d}) \ , \tag{15}$$

ただし,

$$\begin{aligned}
q(V_{d,k} \mid \alpha_{d,k,1}, \alpha_{d,k,2}) &= \frac{\Gamma(\alpha_{k,1} + \alpha_{d,k,2})}{\Gamma(\alpha_{d,k,1})\Gamma(\alpha_{d,k,2})} V_{d,k}^{\alpha_{d,k,1}-1} (1 - V_{d,k})^{\alpha_{d,k,2}-1} \ , \\
q(Z_{d,t} \mid \phi_{d,k}) &= \mathrm{Multinomial}(Z_{d,1}, \cdots, Z_{d,T} \mid 1, \phi_{d,1}, \cdots, \phi_{d,K}) \ .
\end{aligned} \tag{16}$$

のように平均場近似[10)] を用いて E-step が各パラメータ $\mathbf{Z}, \mathbf{C}, \mathbf{V}$ ごとに別々に更新出来る形に置き換える. この近似によって, 式 (14) の E-step における等号成立は必ずしも実現出来なくなるため, 対数尤度が毎反復ごとに単調増加する保証はなくなる. ここで, $\mathbf{1}[\cdot]$ を $\cdot$ が真のときのみ 1 で, 偽のとき 0 とする. このとき,

$$\begin{aligned}
\log p(\mathbf{C} \mid \mathbf{H}, \mathbf{U}, \mathbf{Z}, \Theta) &= -\sum_{\omega,t,d} H_{\omega,d}^{(Z_{d,t})} U_{d,t} + \sum_{\omega,t,d} C_{\omega,t,d} \log H_{\omega,d}^{(Z_{d,t})} U_{d,t} + \tau \\
&= -\sum_{\omega,t,d,k} H_{\omega,d}^{(k)} \mathbf{1}[Z_{d,t} = k] U_{d,t} + \sum_{\omega,t,d,k} C_{\omega,t,d} \mathbf{1}[Z_{d,t} = k] \log H_{\omega,d}^{(k)} U_{d,t} + \tau
\end{aligned} \tag{17}$$

と表せる. ただし, $\tau = \sum_{\omega,t,d} \log \Gamma(C_{\omega,t,d} + 1)$. したがって $q(\mathbf{C})$ の更新は

$$q(\mathbf{C}) \propto \mathrm{Multinominal}(C_{\omega,t,1} \cdots C_{\omega,t,D} \mid Y_{\omega,t}, \lambda_{\omega,t,1}, \cdots, \lambda_{\omega,t,D}) \ , \tag{18}$$

となる. ただし, $r_{d,t}^{(k)} = E_{q(\mathbf{Z})}[\mathbf{1}[Z_{d,t} = k]]$ とし,

$$\lambda_{\omega,t,d} = \frac{\exp\left(\sum_k r_{d,t}^{(k)} \log H_{\omega,d}^{(k)} U_{d,t}\right)}{\sum_d \exp\left(\sum_k r_{d,t}^{(k)} \log H_{\omega,d}^{(k)} U_{d,t}\right)} \ . \tag{19}$$

とする. また, $q(\mathbf{V})$ と $q(Z_{d,t})$ の更新は次のようになる[10)].

$$\alpha_{d,k,1} = 1 + \sum_{t} \phi_{d,t}^{(k)} \ , \quad \alpha_{d,k,2} = \gamma_d + \sum_{t} \sum_{j=k+1}^{K} \phi_{d,t}^{(k)} \ ,$$

$$\phi_{d,t}^{(k)} \propto \exp\left\{E_{q(\mathbf{C})}[\log p(\mathbf{C} \mid \mathbf{H},\mathbf{U},\mathbf{Z},\Theta)] + E_{q(\mathbf{V})}[\log p(Z_{d,t} \mid \mathbf{V})]\right\}$$
$$= \exp\Bigg(E_{q(\mathbf{V})}[\log V_{d,k}] + \sum_{j=k+1}^{K-1} E_{q(\mathbf{V})}[1-\log V_{d,j}]$$
$$-H_{\omega,d}^{(k)}U_{d,t} + \lambda_{\omega,t,d}Y_{\omega,t}\log H_{\omega,d}^{(k)}U_{d,t}\Bigg) \ , \qquad (20)$$

ただし, Ψ を digamma 関数として, ベータ分布の性質に基づいて

$$E_{q(\mathbf{V})}[\log V_{d,k}] = \Psi(\alpha_{d,k,1}) - \Psi(\alpha_{d,k,1}+\alpha_{d,k,2})$$
$$E_{q(\mathbf{V})}[1-\log V_{d,k}] = \Psi(\alpha_{d,k,2}) - \Psi(\alpha_{d,k,1}+\alpha_{d,k,2}) \ . \qquad (21)$$

である. ここで, 式 (13) の $H_{\omega,d}^{(k)}$ に対する一階微分を 0 にすることによって, $H_{\omega,d}^{(k)}$ の更新式が得られる. 同様に $U_{d,t}$ の更新式が得られ, これらは

$$H_{\omega,d}^{(k)} = \frac{\sum_t r_{\omega,t}^{(k)}\lambda_{\omega,t,d}Y_{\omega,t}}{\sum_t r_{\omega,t}^{(k)}U_{d,t}} \ , \quad U_{d,t} = \frac{\eta_1 + \sqrt{\eta_1^2+4\eta_0\eta_2}}{2\eta_0} \qquad (22)$$

と表せる. ただし,

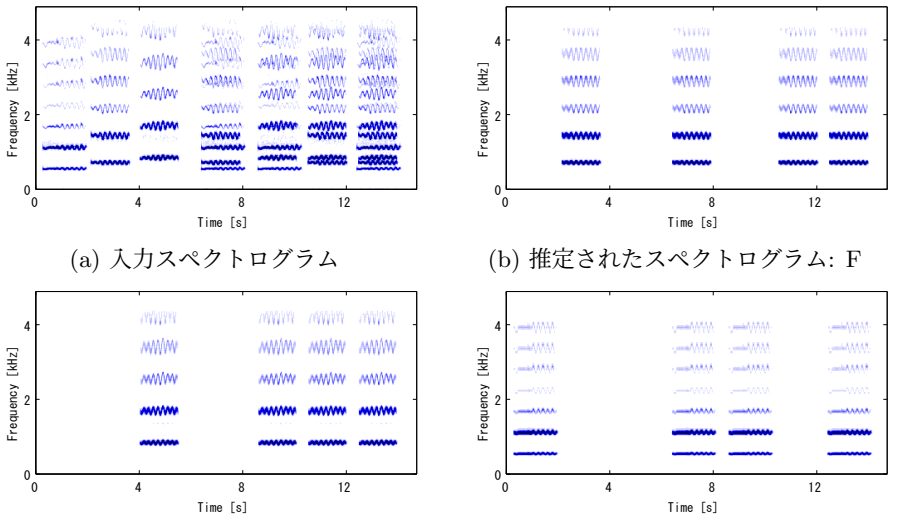
$$\eta_0 = \sum_{\omega,k} H_{\omega,d}^{(k)} r_{d,t}^{(k)} + \frac{\beta_d+1}{U_{d,t+1}} \ , \quad \eta_1 = \sum_{\omega,k} C_{\omega,t,d} r_{d,t}^{(k)} - 1 \ , \quad \eta_2 = (\beta_d+1)U_{d,t-1} \ . \ (23)$$

上述のアルゴリズムを音楽信号のスペクトログラムに適用する場合, 実際は $p(\mathbf{C} \mid \mathbf{H},\mathbf{U},\mathbf{Z},\Theta) = \prod_{\omega,t} p(C_{\omega,t,d} \mid H_{\omega,d}^{(k)}, U_{d,t}, Z_{d,t}, \Theta) << p(\mathbf{Z} \mid \Theta)p(\mathbf{V} \mid \Theta)$ となるため, ディリクレ過程の性質を有効に利用できない場合がある. このような問題を解決するために, ここではスパース性や連続性の制約付き NMF[4),17)] と同様に, 重み付けのためのパラメータ W を導入し,

$$\mathcal{J}(\mathbf{H},\mathbf{U},\mathbf{C},\mathbf{Z}) = \log p(\mathbf{C}|\mathbf{H},\mathbf{U},\mathbf{Z}) + \log p(\mathbf{U}) + \frac{\Omega}{W}\left\{\log p(\mathbf{Z} \mid \mathbf{V}) + \log p(\mathbf{V} \mid \Theta)\right\} \ .$$

とする. よって, 式 (20) は次式に置き換える.

$$\phi_{d,t}^{(k)} \propto \exp\Bigg\{E_{q(\mathbf{V})}[\log V_{d,k}] + \sum_{j=k+1}^{K-1} E_{q(\mathbf{V})}[1-\log V_{d,j}]$$
$$+\frac{W}{\Omega}\left(-H_{\omega,d}^{(k)}U_{d,t} + \lambda_{\omega,t,d}Y_{\omega,t}\log H_{\omega,d}^{(k)}U_{d,t}\right)\Bigg\} \ . \qquad (24)$$



(a) 入力スペクトログラム　(b) 推定されたスペクトログラム: F

(c) 推定されたスペクトルグラム: A♭　(d) 推定されたスペクトルグラム: D♭

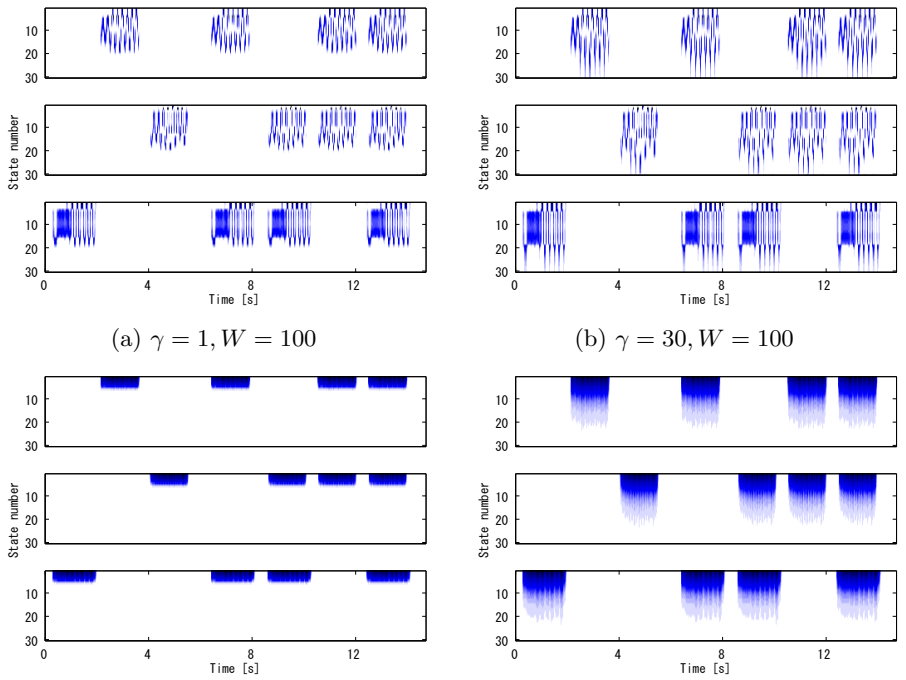
図 2 入力スペクトログラム（RWC-MDB-I-2001 No.45 から生成）(a) と 推定された 3 つのスペクトログラム (b), (c), (d).

4. 実　　験

この章では提案法を音楽信号に適用した実験結果を示す. 振幅スペクトログラムは短時間フーリエ変換 (サンプリング周波数 16kHz, フレーム長 64ms, フレームシフト 32ms, Hanning 窓) により計算した.

まず, RWC-MDB-I-2001 No.45[18)] から生成したボーカルの混合音に提案法を適用した. 信号は D♭, F, A♭ の 3 つの音高で構成されており, はじめに 1 音ごとに演奏し, 次に 2 音ずつの組み合わせを演奏し, 最後に 3 音を同時に発音した. 提案法におけるパラメータは $D=3$, $K'=30$, $\beta_d=0.1, \gamma_d=\gamma=1, W=100$ とした. 図（2）に示す通り, 提案法はボーカルのヴィブラートを学習出来ていることが確認できる.

次に, 無限状態スペクトルに対するパラメータの影響を確認した. 図 (3) に各パラメータにおける $r_{d,t}^{(k)}U_{d,t}$ を示した. 各楽器音を表現するための状態数が自動的に推定されており, さらにそれらはパラメータ γ, W によって調節出来ることが分かる. 小さな γ はより少ない



(a) $\gamma = 1, W = 100$　(b) $\gamma = 30, W = 100$

(c) $\gamma = 1, W = 20$　(d) $\gamma = 30, W = 20$

図 3　$\gamma = 1, W = 100$ (a), $\gamma = 30, W = 100$ (b), $\gamma = 1, W = 20$ (c), $\gamma = 30, W = 20$ (d) と γ, W を変えた時の $r_{d,t}^{(k)} U_{d,t}$ $(d = 1, 2, 3)$ の違い. 上段が F, 中段 A♭, 下段が D♭ を表している.

状態数で信号を表現しようとし, 大きな W はディリクレ過程の効果よりも観測信号をより正確に表現する方を優先するようになることが確認できる.

最後に提案法 (DNMF として参照) を音源分離に用いた場合の性能を評価した. 入力信号は上記の信号に加え, 実際に IC recorder を用いて 5.5m ×3.5m× 3m の部屋で男性ボーカルから録った実信号を用いた. 振幅スペクトログラム上で, 各音源の元信号 $\mathrm{I}_{\omega,\mathrm{t}}$ と推定した $\hat{\mathrm{I}}_{\omega,\mathrm{t}}$ の間の SNR(Signal-to-Noise Ratio) は, $\mathrm{SNR} = 10 \log_{10}(\sum_{\omega,\mathrm{t}} \mathrm{I}_{\omega,\mathrm{t}}^2)/(\sum_{\omega,\mathrm{t}} \left(\mathrm{I}_{\omega,\mathrm{t}} - \hat{\mathrm{I}}_{\omega,\mathrm{t}}\right)^2)$ により計算した. 通常の NMF を比較のために用い, 各アルゴリズムは初期値を乱数で与え 10 回試行し, それらの平均によって評価した. 実信号に対する結果は図 (4) に示した. 音源分離の性能は表 (1) に示した通りで, 提案手法がより音楽のスペクトログラムに適したモデルになっていることが確認できる.

表 1　各アルゴリズムにおける音源分離性能の SNR(dB) とエラー率 (%)
Table 1　Source Separation Performance (SNR [dB])

	RWC database				Real audio data			
Algorithms	S1	S2	S3	Mean	S1	S2	S3	Mean
standard NMF	3.8	3.3	3.5	3.6	7.1	6.9	5.5	6.5
DNMF ($\gamma = 1, W = 20$)	4.2	3.4	3.5	3.7	6.9	8.4	8.2	7.8
DNMF ($\gamma = 30, W = 20$)	4.2	3.8	3.5	3.7	6.9	8.5	8.3	7.9
DNMF ($\gamma = 1, W = 60$)	7.0	10.2	10.4	9.2	10.8	11.2	11.4	11.2
DNMF ($\gamma = 30, W = 60$)	7.1	10.1	10.4	9.2	10.9	11.1	11.2	11.1
DNMF ($\gamma = 1, W = 100$)	9.8	12.4	12.8	11.7	11.3	12.0	11.8	11.7
DNMF ($\gamma = 30, W = 100$)	10.0	12.4	12.8	11.8	11.2	11.7	11.5	11.5

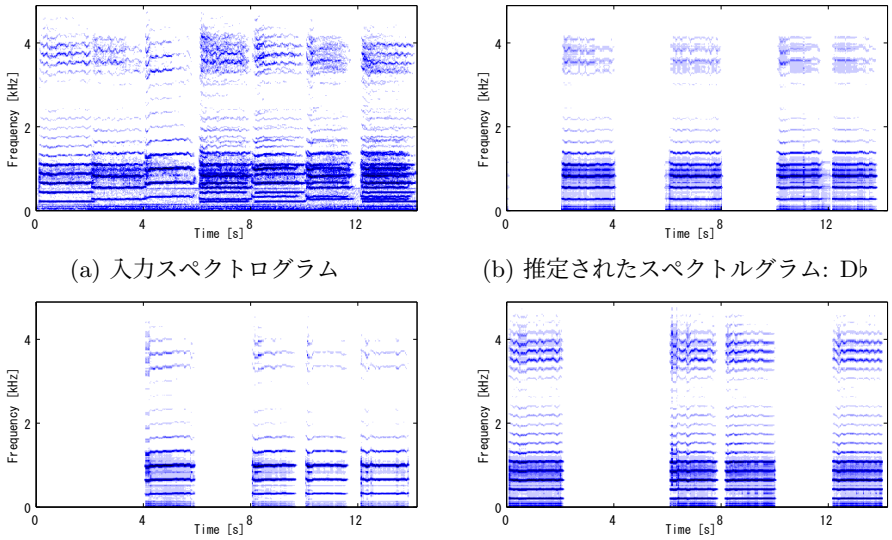
5. おわりに

本報告では, 音楽音響信号解析のための無限状態スペクトルモデルを提案した. パラメータ推定に関しては, 事後確率最大化の観点でのアルゴリズムのみを導出したが, 基底とアクティベーションへの事前分布をガンマ分布にすることによって, 軽微な変更で変分ベイズの観点からのアルゴリズムを導くことができ, またマルコフ連鎖モンテカルロ法などのサンプリングによるパラメータ推定法も可能となる. 今後はディリクレ過程を階層化することによって, 無限状態スペクトルの間の状態遷移を考慮したモデルへ拡張する予定である.

謝辞　本研究の一部は、文部科学省科学研究費補助金基盤研究 (A) (課題番号 00303321)、科学技術振興機構 CrestMuse プロジェクトの支援を受けて行われた。

参考文献

1) D. D. Lee and H. S. Seung, “ Learning the parts of objects by non-negative matrix factorization, ” *Nature*, vol. 401, pp. 788-791, Oct. 1999.
2) D. D. Lee and H. S. Seung, “ Algorithms for non-negative matrix factorization, ” in *Proc. of the Conference on Advances in Neural Information Processing Systems*, vol. 13. Vancouver, British Columbia, Canada: MIT Press, Dec. 2001, pp. 556-562.
3) P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* , Oct. 2003.



(a) 入力スペクトログラム　(b) 推定されたスペクトルグラム: D♭

(c) 推定されたスペクトルグラム: E　(d) 推定されたスペクトルグラム: A

図 4　入力スペクトログラム（実録音）(a) と 推定された 3 つのスペクトログラム (b), (c), (d).

4) T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1066-1074, Mar. 2007.
5) 亀岡, ルルー, 大石, 柏野, "Music Factorizer: 音楽音響信号をノート単位で編集できるインタフェース," 情報処理学会研究報告, 2009-MUS-81-9, 2009.
6) R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non stationary audio events," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 445–448, Mar. 2010
7) 中野, 北野, ルルー, 亀岡, 小野, 嵯峨山, "可変基底 NMF に基づく音楽音響信号の解析," 情報処理学会研究報告, 2009-MUS-81-9, 2009.
8) M. Nakano, J. Le Roux, H. Kameoka, Y. Kitano, N. Ono, S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *9th International Conference on Latent Variable Analysis and Signal Separation*, 2010.
9) A. Ozerov, C. Fevotte and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.
10) D. M. Blei and M. I. Jordan. "Variational inference for Dirichlet process mixtures," *Journal of Bayesian Analysis*, 1(1): pp. 121-144, 2005.
11) J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639-650, 1994.
12) A. T. Cemgil. "Bayesian inference in non-negative matrix factorisation models," Technical Report CUED/F-INFENG/TR.609, University of Cambridge, July 2008.
13) N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. on Audio, Speech and Language Processing*, 18(3), pp. 538-549, 2010.
14) C. Févotte, N. Bertin, and J.-L. Durrieu, " Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis, " *Neural Computation*, vol. 21, no. 3, pp. 793-830, Mar. 2009.
15) C. Févotte and A. T. Cemgil, "Nonnegative matrix factorizations as probabilistic inference in composite models," in *Proc. European Signal Processing Conference*, 2009, vol. 47, pp. 1913-1917.
16) N. Ono, K. Miyamoto, H. Kameoka, S. Sagayama, " A Real-time Equalizer of Harmonic and Percussive Components in Music Signals, "in *Proc. of International Conference on Music Information Retrieval*, pp.139-144, Sep., 2008.
17) P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.
18) M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *Proc. International Conference on Music Information Retrieval*, pp. 287-288, 2002.

付　　録

A.1　確率分布の表記

ポアソン分布: $\mathrm{Poisson}(y \mid x) = \exp(y \log x - x - \log \Gamma(y+1))$

ディリクレ分布: $\mathrm{Dirichlet}(x_1, \cdots, x_K \mid a_1, \cdots, a_K) = (\Gamma(\sum_i a_i) / \prod_i \Gamma(a_i)) \prod_i x_i^{a_i - 1}$

多項分布: $\mathrm{Multinomial}(y_1, \cdots, y_D \mid x, a_1, \cdots, a_D)$

$$= \exp\left(\log \Gamma(x+1) + \sum_d (y_d \log a_d - \log \Gamma(y_d + 1))\right)$$

逆ガンマ分布: $\mathrm{InverseGamma}(y \mid a, b) = \frac{b^a}{\Gamma(a)} y^{-(a+1)} \exp(-b/y)$