社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

［招待講演］階層ベイズ音響・言語モデルに基づく教師なし音楽理解

吉井　和佳†

† 京都大学 大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †yoshii@kuis.kyoto-u.ac.jp

あらまし　本稿では，教師なし音楽理解のための統計的アプローチについて述べる．我々の目標は，言語モデルと音響モデルとを内包する統一的な階層ベイズモデルを定式化することにより，音楽音響信号に対して自動採譜を行う，すなわち音符配置を推定すると同時に，音符配置の背後に存在する音楽文法を同時に推論することである．このアプローチは，音楽信号だけから自己組織的に音響モデル・言語モデルを教師なし学習するという点で，一般的な音声認識システムの枠組みよりは，音声信号からの言語獲得と関連が深い．したがって，音響信号に含まれる音符の個数や音楽文法の複雑さなどを，データに合わせて自動調節できる仕組みが不可欠である．本稿では，音響モデルや言語モデルの一例として，多重基本周波数解析やコード進行解析のためのノンパラメトリックベイズモデルを紹介する．さらに，これらを階層ベイズモデルとして統合する試みについて紹介する．

キーワード　音楽情報処理，多重基本周波数推定，自動採譜，文法獲得，確率的音響モデル，確率的言語モデル，階層ベイズ，ノンパラメトリックベイズ

[Invited Talk] Unsupervised Music Understanding based on Hierarchical Bayesian Acoustic and Language Models

Kazuyoshi YOSHII†

† Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
E-mail: †yoshii@kuis.kyoto-u.ac.jp

Abstract　This paper presents a statistical approach to unsupervised music understanding. Our goal is to estimate musical notes from music audio signals and induce music grammars from the estimated notes by formulating a unified hierarchical Bayesian model consisting of probabilistic acoustic and language models. Given music audio signals, both models are jointly trained in a self-organizing manner. In this paper, we introduce our nonparametric Bayesian acoustic and language models for multipitch analysis and chord progression analysis. We then explain how to unify those models in a hierarchical Bayesian manner.

Key words　Music information processing, multipitch analysis, automatic music transcription, grammar induction, probabilistic acoustic model, probabilistic language model, hierarchical Bayes, nonparametric Bayes

1. はじめに

音楽情報処理分野において，音楽音響信号に対する統計的モデリング（音響モデルの研究）は，ホットなトピックのひとつである．特に，多重音に対する基本周波数推定や音源分離における有用性から，非負値行列分解 (nonnegative matrix factorization: NMF) は大きな注目を集めている [1–4]．標準的な NMF では，多重音の振幅あるいはパワースペクトログラム（非負値行列）を二つの非負値行列，すなわち，周波数方向の基底スペクトルの集合と各基底スペクトルに対応する時間方向の音量変化の集合とに分解することができる．

近年，音声信号の生成過程を説明する目的で考案されたソース・フィルタ理論を楽器音の統計的モデリングに援用することがしばしば行われている [5,6]．周波数領域において，楽器音の音高と音色はそれぞれ，音源信号の性質を表す微細構造（調波構造）と楽器個体の共鳴特性を表すスペクトル包絡によってよく特徴づけられる．人間の聴覚系はスペクトルのピーク（フォルマント）に対して敏感であるため，楽器音の各時間フレームにおけるスペクトル包絡を，全極型周波数伝達関数（自己回帰フィルタの周波数応答）を用いて表現することが一般的である [5]．全極型スペクトル包絡推定の古典的な方法である線形予測分析 (linear predictive coding: LPC) [7] は，音源信号が

ガウス性白色雑音であるという強い仮定のもとで，観測音声信号のスペクトル包絡を推定する手法である．

LPC の問題点を解決する有望な統計的アプローチとして，複合自己回帰モデル (composite autoregressive model: CAR) [2] と呼ばれるソース・フィルタ NMF が提案されている．観測音響信号のスペクトログラムは，複数個の微細構造（ソース）と複数個のスペクトル包絡（フィルタ）との組み合わせから構成されているとみなす．重要なのは，音源信号がガウス性白色雑音であるとは仮定されておらず，全極型スペクトル包絡と同時に音源信号のスペクトル自体が推定される点である．さらに，ノンパラメトリックベイズ拡張された無限複合自己回帰モデル (infinite CAR: iCAR) [3] では，適切な個数のソースとフィルタが推定できるだけではなく，音源信号のスペクトルを調波構造型の関数で表現することで，音源信号の基本周波数 (F0) を推定できる．本稿では，2. 章において，iCAR を対数周波数領域で再定式化した無限重畳離散全極モデル (infinite superimposed discrete all-pole model: iSDAP) [8] を紹介する．

一方，楽譜や和音系列などの離散記号で表現されるデータに対する統計的モデリング（言語モデルの研究）についても近年盛んになりつつある．これまで，和音系列に対する言語モデルとして，n グラムモデルが広く利用されてきた [9–15]．和音は時間局所的な依存性を持つことが知られているため，この仮定はある程度妥当であると考えられている．しかし，従来の経験的なスムージング方法に基づく n グラムモデルを適用する上では，三つの本質的な問題があった．すなわち，1) n グラムモデルの理論的な裏付けがないこと，2) 各和音は異なるコンテキスト長を持っているにもかかわらず，n の値を一意に決める必要があること，3) 語彙として考慮すべき和音の種類や粒度 (例：major, minor, augmented, diminished, seventh, ninth, $\cdots$, それらの派生形) が恣意的に決められること，が問題であった．特に，3) の問題意識を持つ研究はほとんどなく，例えば，和音認識タスクにおいては，語彙に含まれない和音は「未知の和音」としてひとくくりにするか，語彙に含まれる和音のうちで最も近いものとみなすことが一般的であった．

このような問題意識のもと，自然言語処理分野で提案されたノンパラメトリックベイズ n グラムモデル [16–19] を拡張した和音系列に対する語彙フリー無限グラムモデル [20] を 3. 章で紹介する．このモデルでは n グラム確率が階層的にスムージングされており，以下のような利点がある．1) 観測データの背後にある生成モデル（確率過程）を考えることで，次の和音の予測確率を理論的に定式化できる．2) 与えられた和音系列中の各和音が，それぞれ異なる長さのコンテキスト（理論的には無限の長さであってもよい）を持つことを許容する．モデルを学習する過程で，各和音ごとにコンテキスト長の事後分布を推定できる．3) あらゆる音の組合せ（音楽で通常用いられているものに限らない）を和音の種類として許容できる．新しい音の組合せの和音が出現するたびに語彙を必要に応じて拡張していくことができる．副次的な効果として，学習データに含まれる，さまざまな長さを持つ統計的に特徴的な和音進行パターンを見つけることができる．

最後に，4. 章で，近年研究が進みつつある音響モデルと言語モデルとを統合する試み [21] について紹介する．音響モデルに求められる能力は，連続データである音響信号から離散的な記号（音符や和音など）を切り出すことで，言語モデルで取り扱えるようにすることである．一般に，NMF を用いれば，各フレームにおける各音高の音量が推定できるが，音高の存在有無を決定するには閾値処理が必要である．そこで，音響モデルに二値変数を組み込むことで，閾値処理を行うことなく音高の有無を判定すると同時に，その音高の組み合わせの系列の背後に存在する和音系列を推定する方法が提案されている．これは，音楽音響信号だけから和音の概念や和音進行のパターンを獲得しようとするものであり，より高度な音楽文法獲得に向けての有望なアプローチである．

2. 無限重畳離散全極モデル

本章では，連続ウェーブレット変換で得られる対数周波数スペクトログラムに対してソース・フィルタ分解を行うための無限重畳離散全極モデル (iSDAP) を紹介する．本モデルでは，与えられた音楽音響信号に対し，各フレームに含まれる複数の F0（調波構造）を推定すると同時に，複数のスペクトル包絡（楽器の音色）を発見することができる．これを実現するため，スペクトル包絡推定のための離散全極モデル (DAP) [22] を，F0 とスペクトル包絡の同時推定のための複合自己回帰モデル (CAR) [2,3] の枠組みに確率的に統合する．このとき，観測スペクトログラムに合わせてソースとフィルタの個数を自動調節するため，ノンパラメトリックベイズモデルを定式化する．もし，観測データが無限にあれば，理論的には無限個のソースとフィルタが存在するはずである．一方，有限の観測データが与えられた場合は，そこに含まれる高々有限個のソースとフィルタを推定する必要がある．ノンパラメトリックベイズ理論を用いると無限次元の空間内でスパースな学習が可能になる．

2.1 確率モデルの定式化

本節では iSDAP の確率モデルの定式化を行う．いま，対数周波数領域における振幅スペクトログラムを $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ とする．ここで，M は周波数ビンの個数であり，N はフレームの個数である．非負値行列 $\boldsymbol{X}$ を，二つの因子 $\boldsymbol{W}$ および $\boldsymbol{H}$ に分解することを考える．

$$X_{mn} \sim \mathrm{Poisson}\left(\sum_{i=1}^{I\to\infty}\sum_{j=1}^{J\to\infty}\theta_{ni}\phi_j W_{nijm}H_{nij}\right) \tag{1}$$

ここで，θ_{ni} はフレーム n におけるソース i の局所的な重み，ϕ_j は全フレームにおけるフィルタ j の大域的な重みを表す．H_{nij} はフレーム n におけるソース i・フィルタ j の組み合わせの音量を表す．$\{W_{nijm}\}_{m=1}^{M}$ は，フレーム n におけるソース i・フィルタ j から生成されたある楽器音の振幅スペクトルである．

2.1.1 ソース・フィルタの組み合わせ

図 2 で示す通り，振幅スペクトル $\{W_{nijm}\}_{m=1}^{M}$ は対数周波数領域において調波構造を持つことを仮定する．

$$W_{nijm} = \sum_{r=1}^{R} F_{nijr}S_{mnir} \tag{2}$$

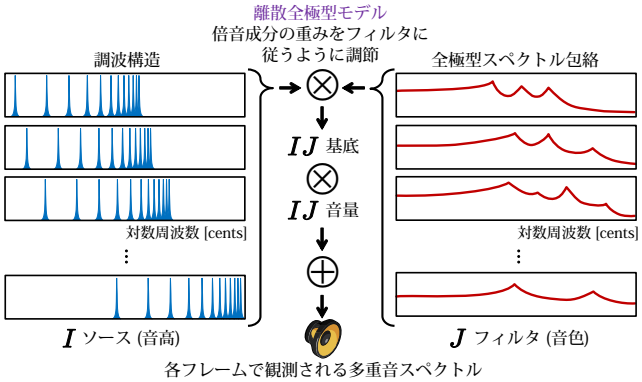

図 1　無限重畳離散全極モデル (iSDAP)：ソース数 I とフィルタ数 J がいずれも無限に発散した場合の極限を考える．

ここで，R は倍音の個数であり，$\{S_{mnir}\}_{m=1}^{M}$ はフレーム n におけるソース i の第 r 次倍音成分の単峰的なスペクトルであり，次式で表すものとする．

$$S_{mnir} = \exp\left(-\frac{1}{2\sigma^2}\left(f_m - (\mu_{ni} + 1200\log_2 r)\right)^2\right) \quad (3)$$

ここで，μ_{ni} はフレーム n におけるソース i の対数基本周波数 F0 [cents]，f_m はスペクトログラムにおける m 番目の周波数ビンに対応する対数周波数，σ^2 は各倍音を中心としたメインローブの広がりである．

ここで，倍音成分の重み $\{F_{nijr}\}_{r=1}^{R}$ は，対数周波数領域において全極型伝達関数を用いて表現する．

$$F_{nijr} = \frac{1}{\left|\sum_{p=0}^{P} a_{jp} e^{-\omega_{nir} p i}\right|} = \left(\boldsymbol{a}_j^T \boldsymbol{U}(\omega_{nir}) \boldsymbol{a}_j\right)^{-\frac{1}{2}} \quad (4)$$

ここで，$\boldsymbol{a}_j \equiv [a_{j0}, \cdots, a_{jP}]^T$ であり，ω_{nir} はフレーム n におけるソース i の第 r 次倍音に対応する正規化角周波数 [rad] であり，$\boldsymbol{U}(\omega)$ は $(P+1)\times(P+1)$ の行列であり，各要素は $[\boldsymbol{U}(\omega)]_{pq} = \cos(\omega(p-q))$ で与えられる．F_{nijr} はパワーではなく，振幅を表すことに注意する．

SDAP では，ソースとフィルタを組み合わせる際に，ソースの倍音成分の重みのみが全極型スペクトル包絡によって制御される．したがって，スペクトル包絡を推定する際に，倍音成分（調波構造のピーク）のみが参照される．これは，CAR は線形予測分析 (LPC) の多重音拡張である野に対して，SDAP が離散全極モデル (DAP) の多重音拡張であることを意味する．

2.2　事前分布の設計

無限次元のベクトル $\boldsymbol{\theta}_n = [\theta_{n1}, \cdots, \theta_{nI}]^T$ および $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_J]^T$ に対してスパースな学習を行うため，ガンマ過程事前分布を仮定する [3, 4]．これを近似的に実現するひとつの方法として，$\boldsymbol{\theta}_n$ および $\boldsymbol{\phi}$ の各要素に対して独立なガンマ事前分布を仮定する．

$$\theta_{ni} \sim \mathrm{Gamma}\left(\frac{\alpha_\theta}{I}, \alpha_\theta\right) \quad (5)$$

$$\phi_j \sim \mathrm{Gamma}\left(\frac{\alpha_\phi}{J}, \alpha_\phi\right) \quad (6)$$

ここで，α_θ および α_ϕ は超パラメータであり，ガンマ過程の集中度と呼ばれる．打ち切りレベル I を無限に大きくしていけば，ベクトル $\boldsymbol{\theta}_n$ は集中度 α_θ をもつガンマ過程からのランダムサンプルとみなすことができる．ここで，任意の正の実数 ϵ に対して $\theta_{ni} > \epsilon$ を満たす実効的な要素数 I^+ はほとんど確実に有限であることが証明されている．現実的には，I を α_θ に比べて十分大きく設定すれば，$\boldsymbol{\theta}_n$ の I 個の要素のいくつかだけがゼロよりある程度大きな値をとることが期待できる．一方，音量 H_{nij} に関しては，ガンマ事前分布を仮定する．

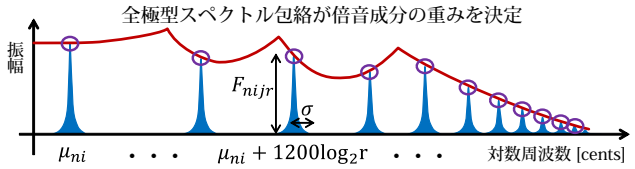

図 2　対数周波数領域におけるフレーム n におけるソース i とフィルタ j との組み合わせによる楽器音スペクトルの生成

$$H_{nij} \sim \mathrm{Gamma}(a_H, b_H) \quad (7)$$

ここで，a_H および b_H は超パラメータである．

3.　語彙フリー無限グラムモデル

本章では，単語系列をモデル化する上で単語の語彙を定義する必要がない NPYLM [19] と同様の問題意識のもと，和音系列をモデル化する上で和音の種類に関する語彙を定義する必要がない語彙フリー無限グラムモデルを紹介する．

3.1　問題設定

いま，語彙（和音の種類の集合）を $\boldsymbol{W}$ とし，そのサイズを V（語彙フリーの場合は音高のあらゆる組み合わせの数）とする．ある和音を $w \in \boldsymbol{W}$ とし，長さ $n-1$ のコンテキストを $\boldsymbol{u} \in \boldsymbol{W}^{n-1}$ とする．学習データ $\boldsymbol{X}$ は，長さ M の和音系列 $x_1 \cdots x_M$ であるとする．ここで，$x_m \in \boldsymbol{W}$ $(1 \leqq m \leqq M)$ であり，簡単のため学習データは 1 つの和音系列のみであるとした．n グラムモデルでは，ある和音 x_m は直前の $n-1$ 個の和音（コンテキスト）に依存していると仮定する．

$$P(x_m | x_1 \cdots x_{m-1}) = P(x_m | x_{m-(n-1)} \cdots x_{m-1}) \quad (8)$$

我々の目的は，学習データ $\boldsymbol{X}$ が与えられたときに，コンテキスト $\boldsymbol{u}$ に続く和音 w の予測確率 $P_{\boldsymbol{u}}(w|\boldsymbol{X})$ を求めることである．これまで，経験的あるいは理論的な様々なスムージング方法が提案されており，以下で説明する．

3.2　階層 Pitman-Yor 言語モデル

Teh [16] は，ノンパラメトリックベイズ理論を用いて，階層 Pitman-Yor 言語モデル (hierarchical Pitman-Yor language model: HPYLM) と呼ぶ新しい n グラムモデルを提案している．興味深いことに，従来より高精度なスムージングができることが経験的に知られていた I-KN は，近年 HPYLM の近似であることが明らかとなった．そのため，より精緻なモデルである HPYLM は，I-KN より優れた予測精度を達成可能である．

3.2.1　階層 Pitman-Yor 過程とモデルの定式化

まず，ノンパラメトリックベイズモデルを構成する際にしばしば利用される Pitman-Yor 過程 (PY) [23] と呼ばれる確率過

程について簡単に説明する．PY はディリクレ過程 (DP) の自由度を増やすように拡張されたものであり，DP と同様に確率分布上の確率分布となっている．すなわち，あるサンプル空間（例：和音の語彙 $\boldsymbol{W}$）上の確率分布（例：あるコンテキストが与えられたときの次の和音の確率分布）に対する確率分布である．したがって，PY は語彙 $\boldsymbol{W}$ 上の n グラム確率分布に対する事前分布として利用可能である．いま，d および θ を正の実数とし，G_0 をサンプル空間上の分布とすると，PY は

$$G \sim \mathrm{PY}(d, \theta, G_0) \tag{9}$$

と書ける．d はディスカウントパラメータ，θ は集中度，G_0 は基底測度と呼ばれる．PY から生成される G もまたサンプル空間上の確率分布である．θ の値が大きくなればなるほど，G_0 と似通った G が生成される確率が高くなる．

HPYLM は PY を階層化することで定式化できる．いま，語彙 $\boldsymbol{W}$ 上のユニグラム分布 G_ϕ があるとする．ここで，ϕ は長さ 0 のコンテキストを表し，$G_\phi(w)$ で和音 w のユニグラム確率を表わすものとする．長さ 1 のコンテキスト $\boldsymbol{u}$ が与えられた上でのバイグラム分布 $G_{\boldsymbol{u}}$ は，ユニグラム分布 G_ϕ とは異なっているものの，（特に高頻度の和音に関して）G_ϕ といくらか類似していると考えられる．このとき，バイグラム分布 $G_{\boldsymbol{u}}$ はユニグラム分布 G_ϕ を基底測度とする PY から生成された，すなわち，$G_{\boldsymbol{u}} \sim \mathrm{PY}(d_1, \theta_1, G_\phi)$ と考える．ここで，d_1 および θ_1 はディスカウントパラメータおよび集中度であり，長さ 1 の任意のコンテキストをもつユニグラム分布の生成過程で共有されている．一般に，長さ $n-1$ のコンテキストが与えられた上での n グラム分布 $G_{\boldsymbol{u}}$ は，$G_{\pi(\boldsymbol{u})}$ を基底測度とする PY から

$$G_{\boldsymbol{u}} \sim \mathrm{PY}(d_{|\boldsymbol{u}|}, \theta_{|\boldsymbol{u}|}, G_{\pi(\boldsymbol{u})}) \tag{10}$$

として生成されると考える．ここで，$d_{|\boldsymbol{u}|}$ および $\theta_{|\boldsymbol{u}|}$ はコンテキスト長 $|\boldsymbol{u}|$ におけるディスカウントパラメータおよび集中度である．ただし，$n-1$ グラム分布 $G_{\pi(\boldsymbol{u})}$ もまた未知であるので，$n-1$ グラム分布 $G_{\pi(\boldsymbol{u})}$ に対する事前分布として，$d_{|\pi(\boldsymbol{u})|}$ および $\theta_{|\pi(\boldsymbol{u})|}$ をパラメータとし，$n-2$ グラム分布 $G_{\pi(\pi(\boldsymbol{u}))}$ を基底測度とする PY を考える．このような生成過程は再帰的に定義することができ，最終的にユニグラム分布 G_ϕ は

$$G_\phi \sim \mathrm{PY}(d_0, \theta_0, G_0) \tag{11}$$

で生成される．ここで，G_0 はグローバルな基底測度であり（ゼログラム分布），通常は一様分布 $G_0(w) = 1/V$ が仮定される．

最終的に，HPYLM は深さ $n-1$ を持つ接尾語木で表現できる．例として，図 3 に $n=3$ の場合を示す．木の中の各ノードごとに，それぞれ対応するコンテキストが存在する．すなわち，木の根ノードから着目するノードまで枝に沿って木を下ることは，コンテキスト中の和音を過去に一つずつ遡ることを意味する．図 3 は中華料理店過程 (CRP) 表現と呼ばれ，ノードはレストラン，ノードに割り付けられた和音 x_m を客と呼ぶ．スムージングのため，客は代理客を上位のレストランに送り込むことがあり，全体の客配置は未知である．逆に言えば，全体の客配置を求めることで，n グラム確率を計算することができる．

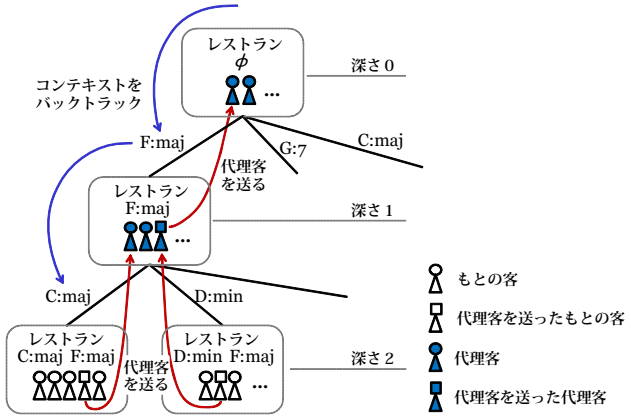

図 3　階層 Pitman-Yor 言語モデルにおける客の着席方法（中華料理店過程: CRP）

3.3　可変長 Pitman-Yor 言語モデル

HPYLM の問題は，全ての客が深さ $n-1$ のレストランに強制的に入れられることである．これは全ての和音のコンテキスト長を $n-1$ に固定していることを意味し，適切ではない．この問題を解決するため，持橋と隅田 [17] は，それぞれの客が異なる深さのレストランに入ることを許容する可変長 Pitman-Yor 言語モデル (variable-order Pitman-Yor language model: VPYLM) を提案している．VPYLM では，それぞれの観測変数 x_m には n グラム長を示す潜在変数 z_m が対応していると仮定する．ただし，z_m の真の値は未知であるので，ベイズ的に z_m の値のあらゆる可能性を考慮に入れて（積分消去することで）次の和音の予測を行う．すなわち，異なる n に対応する n グラムモデルを重みつきで加算することで，無限混合モデルの一種である無限グラムモデルが構成できる．

VPYLM は，客 x_m が入るべきレストランの深さ（あるいは n グラム長 z_m）を確率的にモデル化することで定義できる．一般に，学習データが多くなるほど高次 n グラムモデルを精度よく学習できるようになることから，多く観測されるコンテキストをもつ客ほど，深いレストランに入りやすくなるような機構が必要である．このような要求に合致するものとして，棒折り過程 (stick-breaking process: SBP) [24] を利用する．

いま，潜在変数 z_m の値は SBP に従ってどのように確率的に決まるかについて考える．まず，客 x_m は経路 $\phi \to x_{m-1} \to x_{m-2} \to \cdots$ に沿って木を下ってゆく．その経路において，客が深さ $i-1$ $(1 \leqq i \leqq \infty)$ のレストラン $\boldsymbol{u}_{i-1}$ に到着するたびに，確率 $\eta_{\boldsymbol{u}_{i-1}}$ で停止するか，確率 $1-\eta_{\boldsymbol{u}_{i-1}}$ で通過するかを選択する．したがって，客が木を下っていって深さ $n-1$ のレストランで停止する確率，すなわち $z_m = n$ となる確率は，

$$P_{\boldsymbol{u}}(n|\boldsymbol{\eta}) = \eta_{\boldsymbol{u}_{n-1}} \prod_{i=1}^{n-1} (1 - \eta_{\boldsymbol{u}_{i-1}}) \tag{12}$$

であり，n が大きくなるに従って指数的に減衰する．

いま，木に含まれる各ノード $\boldsymbol{u}$ に対するパラメータ $\eta_{\boldsymbol{u}}$ は未知であるので，事前分布を設定することで不確実性を適切に取り扱うことにする．具体的には，定義域が $0 < \eta_{\boldsymbol{u}} < 1$ であることから，α および β を超パラメータとするベータ事前分布を

設定する．

$$p(\boldsymbol{\eta}) = \prod_{\boldsymbol{u}\in\text{tree}} \text{Beta}(\eta_{\boldsymbol{u}}|\alpha,\beta) \tag{13}$$

3.4 ネスト型 Pitman-Yor 言語モデル

自然言語処理分野における従来の n グラムモデルの本質的な問題は，新語や造語が作られるなどして語彙は日々成長していくものであるにもかかわらず，有限の語彙を明示的に定義しなければならないことであった．単語系列をモデル化する上でこの問題を解決するため，持橋ら [18,19] は可算無限個の単語上に定義された基底測度 G_0 に基づくネスト型 Pitman-Yor 言語モデル (nested Pitman-Yor language model: NPYLM) を提案している．もし，従来通り一様分布 $G_0(w) = 1/V$ を基底測度に用いると，語彙のサイズが $V \to \infty$ であるときに，$G_0(w) \to 0$ となってしまう．代わりに，文字レベルの VPYLM に基づく「綴りモデル」を単語レベルの VPYLM の基底測度に用いることができる．言いかえれば，各単語は文字の系列であり，それらの生成は文字レベルの CRP に従うと仮定する．単語の長さは，非負整数上に定義されたポアソン分布に従うことが知られている．したがって，あらゆる単語 w のゼログラム確率 $G_0(w)$ を，その単語を構成する文字群の生成確率と単語長の確率との積として定義することで，無限語彙モデルを得ることができる．

3.5 音高の組み合わせ系列に対する NPYLM

和音系列をモデル化する上では，単語系列のための n グラムモデルである NPYLM を直接適用することはできない．単語は文字の経時的な並びであるのに対し，和音は音符の同時的な組み合わせであるので，NPYLM とは異なる基底測度 G_0 が必要になる．この問題を解決するため，和音の構成音に基づく確率モデルを和音レベルの VPYLM の基底測度 G_0 として利用する方法を提案する．具体的には，ある和音に 12 種類ある各ピッチクラスの音が「含まれているかどうか」に関するモデル化を行う．この場合，語彙サイズ $V = 49153$ は大きいながらもいまだ有限であり，従来の有限語彙の n グラムモデルでも扱うことはできる．しかし，本研究で提案する方式では，楽譜情報が利用可能であれば，ある和音に各ピッチクラスの音が「何個含まれているか」を非負整数上に定義される分布（ポアソン分布や負の二項分布など）を用いてモデル化することで，容易に無限語彙モデルを定式化できる利点がある．

基底測度 G_0 は，根音のクラスと各構成音の有無とが独立であるという仮定に基づいてモデル化する．一般に和音 w は，$w_0{:}w_1\cdots w_{12}$ と書くことができる．ここで，w_0 は根音のクラスを示す確率変数であり，他はバイナリ値をとる確率変数である．ただし，$w = \text{N}$ であれば，$w_0 = \text{N}$ として残りの変数は利用しないものとする．このとき，w_0 は 13 次元の離散分布に，残りの各変数はそれぞれベルヌイ分布に従うと仮定すると，

$$G_0(w) = p(w|\boldsymbol{\pi},\boldsymbol{\tau}) = \pi_{w_0}\prod_{i=1}^{12}\tau_i^{w_i}(1-\tau_i)^{1-w_i} \tag{14}$$

とできる．ここで，$\boldsymbol{\pi} = \{\pi_{\text{C}}, \pi_{\text{C}\#}, \cdots, \pi_{\text{B}}, \pi_{\text{N}}\}$ は，13 種類のシンボル，すなわち 12 種類の根音のクラスか特別なクラス N の生起する確率であり，$\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_{12}\}$ は対応する相対音高の存在確率を示している．ただし，$w = \text{N}$ であれば，$G_0(w) = \pi_{\text{N}}$ とする．さらに，パラメータ $\boldsymbol{\pi}$ および $\boldsymbol{\tau}$ の値は未知であるので，共役事前分布としてディリクレ分布およびベータ分布を設定することで，不確実性を適切に取り扱う．

$$p(\boldsymbol{\pi},\boldsymbol{\tau}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{a}_0)\prod_{i=1}^{12}\text{Beta}(\tau_i|b_0,c_0) \tag{15}$$

ここで，13 次元ベクトル $\boldsymbol{a}_0$, 正の実数 b_0 および c_0 は超パラメータである．

4. 階層ベイズ音響・言語モデル

本章では，NMF に基づく音響モデルと，HMM に基づく言語モデルを統合することにより，和音と音高の依存関係を考慮しつつ，音楽音響信号に対する音高推定を行うだけではなく，和音進行を学習することができる手法 [21] を紹介する．

4.1 音響モデルの定式化

音響モデルは，通常の NMF に対して，基底のオン・オフを表す二値変数を導入することで定式化できる．

$$X_{mn} \sim \text{Poisson}\left(\sum_{k=1}^{K} W_{km}H_{kn}S_{kn}\right) \tag{16}$$

ここで，$\{W_{mk}\}_{m=1}^{M}$ は k 番目の基底スペクトルを，H_{kn} は時刻 n における基底 k の音量を，S_{kn} は時刻 n において k 番目の基底が使われているかどうかを示す二値変数を表す．

基底スペクトル行列 $\boldsymbol{W}$ に対しては，ガンマ事前分布を仮定し，スパースになるように誘導する．

$$W_{km} \sim \text{Gamma}\,(a_0,b_0) \tag{17}$$

アクティベーション行列 $\boldsymbol{H}$ も基底行列 $\boldsymbol{W}$ と同様にモデル化できる．ただし，H_{kn} がほぼ 0 となってしまうと S_{kn} の値が NMF に影響を与えず，マスクとしての機能を果たさない．そこで，H_{kn} の事前分布として逆ガンマ分布を仮定すれば，H_{kn} が常にある程度の値を持つように誘導すればこの問題は回避できる．さらに，時間方向の滑らかさを導入するため，$\boldsymbol{H}$ に対し下式に示す逆ガンマ連鎖事前分布を与える．

$$G_{kn}|H_{k,n-1} \sim \text{InverseGamma}\left(\eta, \tfrac{\eta}{H_{k,n-1}}\right) \tag{18}$$

$$H_{kn}|G_{kn} \sim \text{InverseGamma}\left(\eta, \tfrac{\eta}{G_{kn}}\right) \tag{19}$$

ここで，η はアクティベーションの時間変化の滑らかさを決定するハイパーパラメータで，G_{kt} は $H_{k,n-1}$ と H_{kn} に正の相関を持たせるために持導入した補助変数である．

4.2 言語モデルの定式化

言語モデルは，マルコフ性を持つ和音列 $\boldsymbol{Z} = \{z_1,\ldots,z_N\}$ $(z_t \in \{1,\ldots,I\})$ を隠れ変数に持ち，二値変数 $\boldsymbol{S} = \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N\}$ $(\boldsymbol{s}_t \in \{0,1\}^K)$ を出力する HMM として定式化される（図 4）．ここで I は隠れ状態の種類，すなわち和音の種類であり，K は出現する可能性がある音高の数を表す．ただし，提案モデル全体から見た場合は，音高の有無を表す $\boldsymbol{S}$ は隠れ変数である．HMM は以下に示すように定式化される．

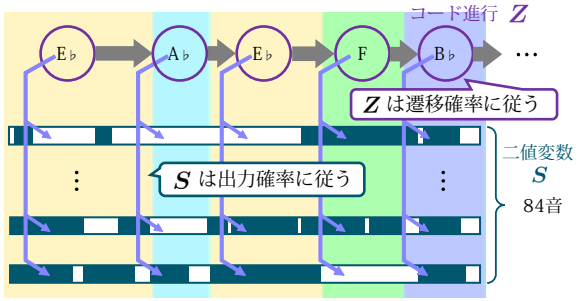

図 4　統一的階層ベイズモデルにおける言語モデル部分.

$$z_1 \sim \mathrm{Categorial}(\boldsymbol{\phi}) \tag{20}$$

$$z_n | z_{n-1} \sim \mathrm{Categorical}(\boldsymbol{\psi}_{z_{n-1}}) \tag{21}$$

$$S_{kn} | z_n \sim \mathrm{Bernoulli}(\pi_{z_n k}) \tag{22}$$

ここで，$\boldsymbol{\psi}_i \in \mathbb{R}^I$ は和音 z_i から次の和音への遷移確率，$\boldsymbol{\phi} \in \mathbb{R}^I$ は初期確率，$\pi_{z_t k}$ は和音 z_t の下で k 番目の音高が出力される確率を表す．これらのパラメータに対し，共役事前分布をおく．

$$\boldsymbol{\psi}_i \sim \mathrm{Dir}(\mathbf{1}_I) \quad \boldsymbol{\phi} \sim \mathrm{Dir}(\mathbf{1}_I) \quad \pi_{ik} \sim \mathrm{Beta}(e, f) \tag{23}$$

ここで，e と f はハイパーパラメータである．

4.3　事後分布の推論

観測データ $\boldsymbol{X}$ が与えられた時に，すべてのパラメータと潜在変数の同時的な事後分布を，ギブスサンプリングを用いて近似的に求めることができる．もし，音高配置を表す二値変数 $\boldsymbol{S}$ が既知であれば，音響モデルと言語モデルは独立に更新できる．一方，両方のモデルが既知であれば，二値変数 $\boldsymbol{S}$ を更新できる．したがって，これらのステップを交互に反復することを繰り返せば，事後分布からのサンプルが近似的に得られる．

5.　おわりに

本稿では，言語モデルと音響モデルとを内包する統一的な階層ベイズモデルを定式化することにより，音楽音響信号に対して自動採譜を行う，すなわち音符配置を推定すると同時に，音符配置の背後に存在する音楽文法を同時に推論する試みについて紹介した．4. 章で述べたように，NMF と HMM（バイグラムモデル）を統合する方式においては，各部について 2. 章や 3. 章で述べた拡張が可能である．一方，音符配置をより精緻にモデル化するためには，言語モデルの高度化が必須である．具体的には，自然言語処理における構文解析や意味解析に相当する処理が必要であり，確率的文脈自由文法 (PCFG) や潜在的ディリクレ配分法 (LDA) などの応用が有望である．

謝辞: 本研究の一部は，JSPS 科研費 24220006, 26700020, 26280089, 16H01744, JST CREST OngaCREST，および栢森情報科学振興財団の支援を受けた．

文　献

[1] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman. Dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, 2014.

[2] H. Kameoka and K. Kashino. Composite autoregressive system for sparse source-filter representation of speech. In *ISCAS*, pp. 2477–2480, 2009.

[3] K. Yoshii and M. Goto. Infinite composite autoregressive models for music signal analysis. In *ISMIR*, pp. 79–84, 2012.

[4] M. Hoffman, D. Blei, and P. Cook. Bayesian nonparametric matrix factorization for recorded music. In *ICML*, pp. 439–446, 2010.

[5] R. Hennequin, R. Badeau, and B. David. NMF with time-frequency activations to model nonstationary audio events. *IEEE trans. on Audio, Speech, and Language Processing*, 19(4):744–753, 2011.

[6] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE trans. on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

[7] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *ICA*, pp. C17–C20, 1968.

[8] K. Yoshii and M. Goto. Infinite superimposed discrete all-pole modeling for source-filter decomposition of wavelet spectrograms. In *ISMIR*, pp. 86–92, 2014.

[9] R. Scholz, E. Vincent, and F. Bimbot. Robust modeling of musical chord sequences using probabilistic N-grams. In *ICASSP*, pp. 53–56, 2009.

[10] 金子 仁美, 川上 大輔, 嵯峨山 茂樹. 機能和声解析データの作成とその統計解析. **音楽情報科学研究会 研究報告**, 2010-MUS-85, pp. 1–8. 情報処理学会, 2010.

[11] M. Ogihara and T. Li. N-gram chord profiles for composer style representation. In *ISMIR*, pp. 671–676, 2008.

[12] C. Pérez-Sancho, D. Rizo, and J. M. I nesta. Genre classification using chords and stochastic language models. *Connection Science*, 21(2-3):145–159, 2009.

[13] H.-T. Cheng, Y.-H. Yang, Y.-C. Lin, I-B. Liao, and H. H. Chen. Automatic chord recognition for music classification and retrieval. In *ICME*, pp. 1505–1508, 2008.

[14] 須見 康平, 糸山 克寿, 吉井 和佳, 駒谷 和範, 尾形 哲也, 奥乃 博. ベース音高と和音特徴の統合に基づく和音系列認識. **情報処理学会論文誌**, 52(4):1803–1812, 2011.

[15] M. Khadkevich and M. Omologo. Use of hidden Markov models and factored language models for automatic chord recognition. In *ISMIR*, pp. 561–566, 2009.

[16] Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, NUS School of Computing, 2006.

[17] 持橋 大地, 隅田 英一郎. 階層 Pitman-Yor 過程に基づく可変長 n-gram 言語モデル. **情報処理学会論文誌**, 48(12):4023–4032, 2007.

[18] 持橋 大地, 山田 武士, 上田 修功. ベイズ階層言語モデルによる教師なし形態素解析. **自然言語処理研究会 研究報告**, 2009-NL-190, 情報処理学会, 2009.

[19] D. Mochihashi, T. Yamada, and N. Ueda. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *ACL-IJCNLP*, pp. 1000–1008, 2009.

[20] K. Yoshii and M. Goto. A vocabulary-free infinity-gram model for nonparametric Bayesian chord progression analysis. In *ISMIR*, 2011. submitted for publication.

[21] Y. Ojima, K. Itoyama, and K. Yoshii. A hierarchical Bayesian model of chords, pitches, and spectrograms for multipitch analysis. In *ISMIR*, pp. 309–315, 2016.

[22] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE trans. on Signal Processing*, 39(2):411–423, 1991.

[23] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[24] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.