3Q-06

コード進行と多重音スペクトルの階層ベイズモデルに基づく音楽音響信号の音高推定

尾島 優太 † 中村 栄太 ‡ 糸山 克寿 ‡ 吉井 和佳 ‡
† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

自動採譜とは，音楽音響信号中に含まれる音高と音価を推定する問題であり，特に音高推定については多くの研究がなされている．従来，音高推定を実現するには，混合音である音楽音響信号の音源分離が必要であると考えられてきた．例えば，非負値行列因子分解 (NMF) を用いると，混合音のスペクトログラムから，異なる音高に対応する基底スペクトルと各基底のアクティベーション (音量) を推定できる [1,2]．ここで，アクティベーションは連続量であるので，各時刻における各音高の on/off を判定 (音高推定＝ピアノロールの出力) するためには，対応する基底スペクトルのアクティベーションに対する二値化を行う必要があった．しかし，閾値設定が困難であるだけではなく，高精度な音源分離ができれば音高推定も容易になるのと同様，音高が既知であれば音源分離は容易になるという鶏と卵の関係が存在するため，このような縦列処理は本来適切ではなかった．

本稿では，音源分離と音高推定を一挙に行うため，混合音の生成過程を確率的に定式化し，その逆問題を解くというアプローチをとる [1,3] (図 1)．まず，ベータ過程 NMF (BP-NMF) [4] と同様，各時刻において各基底の on/off を制御する二値変数を導入する．これにより，アクティベーションの値に関わらず，二値変数が off を取る場合には，対応する音高は混合音の生成には寄与しなくなる．さらに，各時刻における音高の組み合わせはコードに依存することから，コード遷移を潜在変数系列とし，音高群の on/off (ピアノロール) を出力する隠れマルコフモデル (HMM) を構成する．すなわち，HMM を音高群に対する事前分布，NMF を混合音に対する音高群の尤度関数とした統一的な階層ベイズモデルが定式化される．この種の言語モデルと音響モデルの統合は，一般的な音声認識システムと同様であるが，ギブスサンプリングを用いて音高推定を行うと同時に，両モデルを一挙に教師なし学習する点が異なる．

2. 提案法

提案モデルは，音響モデルである NMF と言語モデルである HMM からなる．以下に確率モデルの定式化とパラメータの事後分布の推定方法について説明する．

2.1 音響モデルの定式化

二値変数が導入された NMF では，混合音の振幅スペクトログラム $\boldsymbol{X} \in \mathbb{R}_+^{F\times T}$ に対して，$\boldsymbol{X} \approx \boldsymbol{W}(\boldsymbol{H}\odot\boldsymbol{S})$ という低ランク近似を行う．ここで，$\boldsymbol{W} \in \mathbb{R}_+^{F\times K}$ は K 個の基底スペクトル群，$\boldsymbol{H} \in \mathbb{R}_+^{K\times T}$ および $\boldsymbol{S} \in \{0,1\}^{K\times T}$ は対応するアクティベーションベクトル群および二値変数ベクトル群であり，F は周波数ビン数，T はフレーム数を表す．また，周波数ビンを f $(1 \le f \le F)$，フレームを t $(1 \le t \le T)$，基底を k $(1 \le k \le K)$ で表す．

NMF は観測行列 $\boldsymbol{X}$ と再構成行列 $\boldsymbol{W}(\boldsymbol{H}\odot\boldsymbol{S})$ との近似誤差を最小化する問題であるが，次式で定義される尤度関数最大化としての解釈が可能である．

$$X_{ft} \mid W_{fk}H_{kt}S_{kt} \sim \mathrm{Poisson}\left(\textstyle\sum_k W_{fk}H_{kt}S_{kt}\right)$$

ここで，尤度関数としてポアソン分布を用いた [1]．本研究ではさらに，事前分布を導入することにより，NMF のベイズモデルを構成する．

$$W_{fk} \sim \mathrm{Gamma}\,(a,b)\,,\ H_{kt} \sim \mathrm{InverseGamma}\,(c,d)$$

ここで，$\boldsymbol{W}$ にはガンマ事前分布を与えることで [5]，調波構造のようなスパースな基底スペクトルに誘導した．一方，$\boldsymbol{H}$ には逆ガンマ分布を与えることで，非スパースなアクティベーションに誘導した．H_{kt} がほぼ 0 である場合には，二値変数 S_{kt} の値に関わらずその基底 k は尤度関数に寄与しないため，S_{kt} の推定が適切に行えなくなってしまう．一方，H_{kt} が常にある程度の大きさをもつようであれば，実際にはその音高が存在しなければ，S_{kt} は 0 を取らざるを得ず，適切にマスクとしての機能を果たす．実際には，時間方向の滑らかさを考慮するため，逆ガンマ連鎖事前分布を用いた [5]．

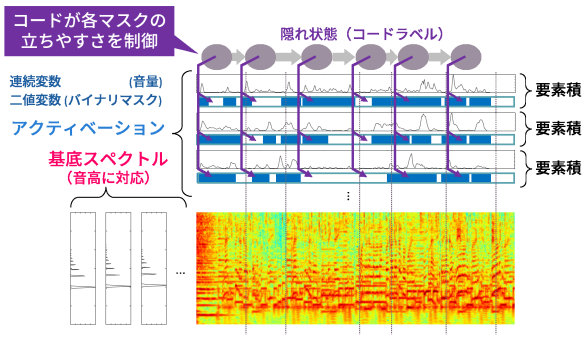

図 1: モデルの全体像

2.2 言語モデルの定式化

コード遷移を潜在変数系列 $\boldsymbol{Z} = \{z_1, \cdots, z_T\}$ $(z_t \in \{1, \cdots, I\})$ に持ち，観測変数系列 $\boldsymbol{S} = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_T\}$ (提案モデル全体から見れば潜在変数) を出力するベイズ HMM を定式化する．ここで，I はコードの種類数 (状態数) とした．まず，状態遷移モデルは次式で与えられる．

$$z_t \mid z_{t-1}, \psi_{z_{t-1}} \sim \mathrm{Categorical}(\boldsymbol{\psi}_{z_{t-1}}), \boldsymbol{\psi}_i \sim \mathrm{Dirichlet}(\mathbf{1}_I)$$

ここで，$\boldsymbol{\psi}_i$ はコード i における遷移確率であり，無情報ディリクレ事前分布を与えた．一方，音高群の出力モデルは次式で与えられる．

$$S_{kt} \mid z_t, \pi_{z_t k} \sim \mathrm{Bernoulli}(\pi_{z_t k}),\ \pi_{z_t k} \sim \mathrm{Beta}(e,f)$$

ここで，各時刻 t における音高 k の on/off は表が出る確率 $\pi_{z_t k}$ のコイントスで決定される．すなわち，$S_{kt} = 1$ となる確率は，その時刻におけるコードによって異なる．実際には，構成音の音高の相対的な関係が同じになるコード同士は，音高の出現確率を共有するようにした．

2.3 事後分布の推論

事後分布 $p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\psi} | \boldsymbol{X})$ を近似計算するため，ギブスサンプリングを用いる．まず，NMF のパラメータ $\boldsymbol{W}, \boldsymbol{H}$ と HMM のパラメータ $\boldsymbol{\pi}, \boldsymbol{\psi}$ を適当に初期化し，$p(\boldsymbol{S} | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{X})$ から音高群 $\boldsymbol{S}$ をサンプルす

Pitch Estimation for Music Audio Signals based on a Hierarchical Bayesian Model of Chords and Spectrograms: *Yuta Ojima, Eita Nakamura, Katsutoshi Itoyama, and Kazuyoshi Yoshii (Kyoto Univ.)*

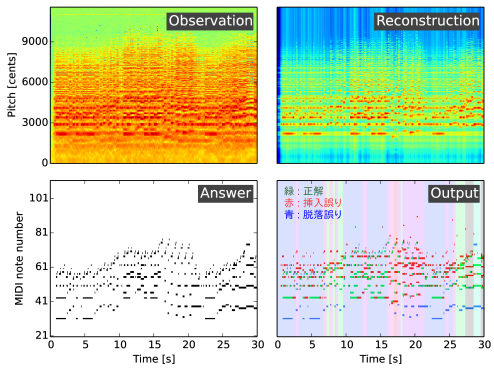

図 2: "MUS-scn15_12_ENSTDkCl"に対する音高推定結果．音高推定結果の背景色はコード推定結果を表す．

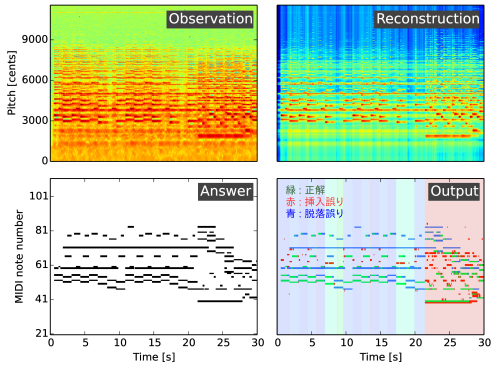

図 3: "MUS-scn16_3_ENSTDkCl"に対する音高推定結果

る．次に，$p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{S},\boldsymbol{X})$ から $\boldsymbol{W},\boldsymbol{H}$ を，$p(\boldsymbol{\pi},\boldsymbol{\psi},\boldsymbol{Z}|\boldsymbol{S})$ から $\boldsymbol{\pi},\boldsymbol{\psi},\boldsymbol{Z}$ をそれぞれサンプルする．このとき，NMF と HMM は文献 [4,6] を参考に独立に更新できる．これを反復することで事後分布に収束する．このように，音響モデルと言語モデルを一挙に最適化して音高推定を行うアプローチは，文献 [7] でも見られる．

3. 実験

本手法の有効性を確認するため，MAPS データベース [8] のうち"ENSTDkCl"のラベルが付されたピアノ曲 30 曲の冒頭 30 秒間に対し，音高推定を行い，基底の制限のみを加えた BP-NMF と比較した．入力音響信号として予め variable-Q 変換 [9] およびリサンプリングにより，周波数ビン 926，時間フレーム 3000 のスペクトログラムへと変換した後，調波打楽器音分離 [10] を行ったものを用いた．なお，ハイパーパラメータは実験的に決定した．また，HMM の初期確率の事前分布は一様であるとし，遷移確率は，フレーム単位でのコードの遷移は起こりにくいため自己遷移を 1–5.0×10^{-8} とし，他状態への遷移がディリクレ分布に従うと仮定した．コードは 12 種類のルート音に対し，Major と Minor の 2 つのコード形が学習されることを意図して計 24 種類を用意した．基底は調波構造を表す基底と非調波構造である雑音を表現する基底を 1 つずつ用意した．調波構造基底をシフトすることで音高に対応させる．また，雑音表現基底のスパース性は言語モデルに依存させず，従来の BP-NMF と同様に決定した．評価は音高推定結果と正解データを基に算出した f 値により行う．なお，評価にあたって曲全体のオクターブの誤りは許容した．

評価実験の結果を表 1 に示す．表 1 より，言語モデルを加えても音高推定精度の平均値は向上しなかった．しかし，曲ごとに結果を比較すると 30 曲中 18 曲に対する音高推定精度が向上していた．このことから，一部の曲

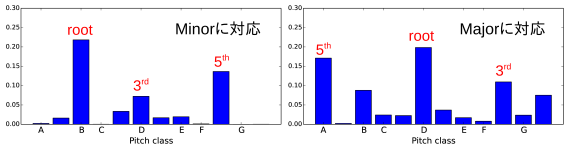

図 4: "MUS-scn15_12_ENSTDkCl"におけるコード構造推定結果

表 1: 音高推定結果の f 値

モデル	f 値	再現率	適合率
言語 + 音響モデル	63.44%	70.02%	58.00%
音響モデルのみ	64.05%	73.82%	56.56%

に対して大きく精度が落ちていることがわかる．

言語モデルの統合により音高推定精度が 3.3 ポイント向上した曲を図 2 に，7 ポイント悪化した曲を図 3 に示す．また，前者に対するコード構造推定結果を図 4 に示す．図 2,3 より，同じ音の響きを持つ繰り返し区間は同一の状態が割り当てられており，言語モデルが正しく学習されていることが分かる．また，図 4 より，事前情報がなくとも音楽音響信号からコードの構造が概ね正しく獲得できていることが分かる．図 2 では残響音を拾う誤りが，図 3 では頻出する音がコード構造として正しく学習できないことによる誤りが見られる．

4. おわりに

本稿では HMM による言語モデルと BP-NMF による音響モデルを統合し，音高推定を行うための手法を提案した．コード構造を考慮しても音高推定結果は向上しなかったが，これは曲によりコード構造を獲得できなかったことが原因であると考えられる．そのため言語モデルに改善を加えることでコード構造を正しく獲得できれば，精度の向上が期待できる．これを踏まえ，今後は隠れセミマルコフモデルを導入することでコード進行をフレーム単位ではなく拍単位で捉え，コード構造推定結果の精度向上を図る予定である．

謝辞 本研究の一部は JSPS 科研費 24220006, 26700020, 26280089, JST CREST の支援を受けた．

参考文献

[1] P. Smaragdis *et al.* Non-negative matrix factorization for polyphonic music transcription. *IEEE WASPAA*, 177–180, 2003.

[2] N. Bertin *et al.* Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE TASLP*, 18(3):538–549, 2010.

[3] K. Yoshii *et al.* Infinite composite autoregressive models for music signal analysis. *ISMIR*, 79–84, 2012.

[4] D. Liang *et al.* Beta process non-negative matrix factorization with stochastic structured mean-field variational inference. *arXiv*, 1411.1804, 2014.

[5] A T. Cemgil *et al.* Conjugate gamma markov random fields for modelling nonstationary sources. *ICA*, 697–705. Springer, 2007.

[6] S. L. Scott. Bayesian methods for hidden Markov models. *JASA*, 97(457), 2002.

[7] H. Kameoka *et al.* Context-free 2D tree structure model of musical notes for Bayesian modeling of polyphonic spectrograms. *ISMIR*, 307–312, 2012.

[8] V. Emiya *et al.* Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE TASLP*, 18(6):1643–1654, 2010.

[9] C. Schörkhuber *et al.* A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution. *AES Conf. SA*, 2014.

[10] D. Fitzgerald. Harmonic/percussive separation using median filtering. *DAFX*, 1–4, 2010.