コード制約付き NMF を用いた音高推定に基づくコード認識

丸尾 智志[1]　吉井 和佳[1]　糸山 克寿[1]　Matthias Mauch[2]　後藤 真孝[3]
[1]京都大学 大学院情報学研究科 知能情報学専攻 [2]Queen Mary University of London [3]産業技術総合研究所

1. はじめに

音楽音響信号に対する自動コード認識は，音楽情報処理の分野における基本的な課題の一つである．コードパターンは音楽の雰囲気に密接に関わっており，ジャンル分類 [1] や楽曲推薦 [2] などに利用できることがその理由である．

コード認識の従来法は一般的に音響特徴量ベクトルの抽出とその分類で構成される．特徴量抽出の最も一般的な方法は，半拍のような短いフレームごとに，C，C#，…，B の 12 のピッチクラスのエネルギーの分布を表現する 12 次元クロマベクトルを計算するものである [3]．一方，特徴量分類には，コードの遷移確率とコードの種類ごとのクロマベクトルの出力確率を表した隠れマルコフモデル（HMM）がよく用いられる [4]．

コードとその構成音には強い関連性があることから，それを利用したコード認識 [5] や音高推定 [6] の精度向上に関する研究が行われている．しかし，それらはコード認識と音高推定の一方向の依存のみを扱っている．

本稿では，コード認識と音高推定の相互依存性に焦点を当てた，コード認識手法（図 1）を提案する．まず，Bayesian NMF により入力の音楽音響信号の音高推定を行い，その結果からクロマベクトルを計算し，Bayesian HMM によるコード認識を行う．次に，その結果を用いて Bayesian NMF に事前分布の形でコード制約を与え，もう一度音高推定およびコード認識を行う．これによりクロマベクトルは倍音の影響が低減されたより鮮明なものになり，コード認識率の向上が期待できる．

2. 音高推定に基づくコード認識

ここでは，提案手法であるコード制約付き NMF を用いた音高推定に基づくコード認識について述べる．本手法は，Bayesian HMM によるコード認識（2.2 節）と Bayesian NMF による音高推定（2.3 節）で構成される．

2.1 問題設定

コード認識は，音楽音響信号からコードラベルの系列を得る問題である．すなわち，対象の信号から得られるクロマベクトルの系列を $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$，それに対応するコードラベルの系列を $\boldsymbol{Z} = \{z_1, z_2, \cdots, z_T\}$ とすると，$\boldsymbol{X}$ から $\boldsymbol{Z}$ を得ることが目標となる．ここで，T は系列の長さである．

本稿では，以下の仮定の下でコード認識を行う．

- 正しい拍の位置は最初から与えられているとする．
- コードの境界は拍の位置（四分音符単位）もしくは半拍の位置（八分音符単位）であるとする．
- コードのタイプは “maj”，“min” および “no chord” とする．他のタイプ（例えば，“maj7” や “dim”）は [7] を参考に “maj” か “min” のどちらかに分類する．
- 調は楽曲の初めから終わりまで同一であるとする（転調は考慮しない）．

Chord Recognition based on Approximate Note Transcription Using NMF with Chord Constraints: *Satoshi Maruo, Kazuyoshi Yoshii, Katsutoshi Itoyama (Kyoto Univ.), Matthias Mauch (Queen Mary University of London), and Masataka Goto (AIST)*

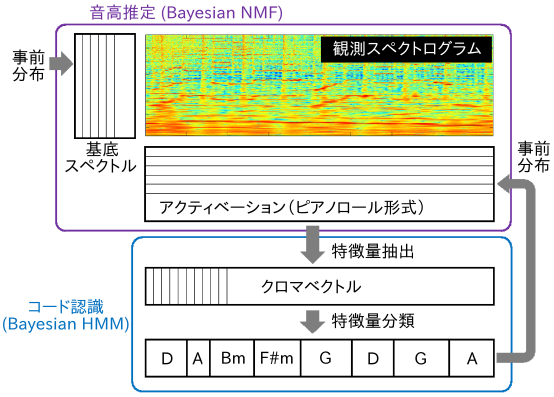

図 1: 提案手法の全体図

2.2 Bayesian HMM によるコード認識

[4] と同様に，24 の調（長調と短調それぞれ 12 個ずつ）に対応する HMM を用いて，コード認識を行う．ただし，本手法では過学習を防ぐために通常の HMM ではなく Bayesian HMM を使用している．

2.2.1 特徴量抽出

本手法では，NMF のアクティベーションからクロマ特徴量を抽出する．これは [5] に類似した方法だが，事前分布の形でコード制約を与えるために，NNLS の代わりに Bayesian NMF を使用する．

各フレーム t において，Bayesian NMF で得られるピッチごとのアクティベーションから 12 次元クロマベクトル $\boldsymbol{x}_t$ を計算する．本稿では，フレームは半拍の区間として扱う．また，扱うピッチは C2 から B6 までの 5 オクターブとする．すなわち，$\boldsymbol{x}_t$ の各次元の値を計算するために，フレーム t において同じピッチクラスに属する 5 つのピッチ（例えば，C2，C3，C4，C5，C6）の振幅値を加算する．最後に，クロマベクトルの各次元の平均が 0，分散が 1 となるように，全てのフレームで標準化する．

2.2.2 学習

HMM の学習は，コード遷移確率の学習と特徴量出力確率の学習に分類される．学習時は，データを最大限使用するために，ピッチクラスの循環性を利用する．

コード遷移確率の学習　それぞれの遷移が発生した回数を数えることで，遷移確率を学習する．全ての調を C に移調し，それに伴ってコードのルート音もシフトさせる．したがって，2 つの調（Cmajor と Cminor）に対してのみコード遷移確率を学習する．他の 22 の調の遷移確率は要素を並び替えることで得られる．

特徴量出力確率の学習　各コード区間における特徴量の出力確率を GMM で学習する．各コードのルート音はクロマベクトルの要素を巡回シフトすることで C に統一する．したがって，2 つのコードタイプ（“maj” と “min”）に対してのみ GMM を学習することになる．他の 22 のコードの出力確率はガウス分布のパラメータを並び替えることで得られる．

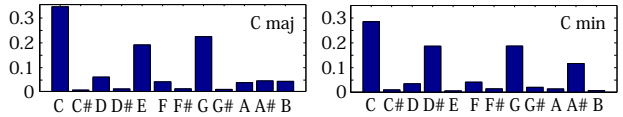

図 2: Cmaj と Cmin の範囲でのピッチクラスの分布

2.2.3 コード認識

学習した HMM を使用し，ビタビアルゴリズムによって最尤経路を探索することで，コード認識を行う．ビタビアルゴリズムによる探索は 24 の調の HMM それぞれに対して別々に行い，尤度が最も高くなる HMM の調を入力の音楽音響信号の調とする．

2.3 Bayesian NMF による音高推定

2.3.1 モデルの定式化

NMF は音源分離や音楽音響信号に対する多重音解析によく用いられている．NMF では，非負の行列（スペクトログラム）$\boldsymbol{Y} \in \mathbb{R}^{M\times N}$ を $\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{H}$ となるように，二つの非負の行列 $\boldsymbol{W} \in \mathbb{R}^{M\times R}$（基底スペクトル）と $\boldsymbol{H} \in \mathbb{R}^{R\times N}$（アクティベーション）に分解する．

Bayesian NMF では，$\boldsymbol{W}$ と $\boldsymbol{H}$ の各要素にガンマ事前分布を与える．事後確率 $p(\boldsymbol{W}, \boldsymbol{H}|\boldsymbol{Y})$ を解析的に求めることはできないため，変分ベイズ法を用いて以下のような近似を行う．

$$p(\boldsymbol{W}, \boldsymbol{H}|\boldsymbol{Y}) \approx q(\boldsymbol{W})q(\boldsymbol{H}) \tag{1}$$

結果として，$q(\boldsymbol{W})$ と $q(\boldsymbol{H})$ は以下の形で得られる．

$$q(w_{rm}) = \mathcal{G}(w_{rm}|a^w_{rm}+\textstyle\sum_n \lambda_{mnk}Y_{mn}, b^w_{rm}+\sum_n \mathbb{E}[h_{rn}]) \tag{2}$$

$$q(h_{rn}) = \mathcal{G}(h_{rn}|a^h_{rn}+\textstyle\sum_m \lambda_{mnk}Y_{mn}, b^h_{rn}+\sum_m \mathbb{E}[w_{rm}]) \tag{3}$$

ただし，$\lambda_{mnk} \propto \exp(\mathbb{E}[\log w_{rm}] + \mathbb{E}[\log h_{rn}])$ は $\sum_k \lambda_{mnk} = 1$ となる補助変数であり，$\mathcal{G}$ はガンマ分布を表す．

2.3.2 ハイパーパラメータの設定

$p(\boldsymbol{W})$ は基底スペクトルが半音階に相当する調波構造をもつように設定する．調波構造において，倍音のパワーは 0.5 倍ずつ指数的に減少すると仮定する．$\boldsymbol{a}^w_r = [a^w_{r1}, \cdots, a^w_{rM}]^T$ は調波構造に比例するように設定する．$\boldsymbol{a}^w_r$ の値を大きくすることで，事前分布の効果を大きくすることができる．

同様に，$p(\boldsymbol{H})$ は各コード区間で特定のピッチクラスが現れやすくなるように設定する．コードの正解データのある MIDI を用いて，図 2 のように Cmaj と Cmin の二つのコードの範囲でどのピッチクラスが現れやすいかを学習する．学習時には，各コード区間の音符はコードのルート音が C となるように置き換える．a^h_{rn} は対応するピッチクラスの確率に比例するように設定する．

3. 実験

3.1 実験条件

The Beatles の楽曲のうち長調の 137 曲を使用し，5 分割の交差検定を行った．長調の楽曲のみを使用するため，12 の長調に対応する 12 の HMM を学習した．音楽音響信号は定 Q 変換により，1 オクターブあたりのビン数が 36，時間間隔が 64 ms のスペクトログラムに変換した．事前分布を使用しない場合は $a^w_{rm} = b^w_{rm} = a^h_{rn} = b^h_{rn} = 1$ とした．コード認識の結果は，正解データと一致する部分の割合で評価した．比較のため，MIR toolbox [8] で計算した標準的なクロマベクトルによるコード認識も行った．

表 1: 実験結果

特徴量	クロマグラム	NMF	制約付き NMF
認識率 (%)	69.0	71.6	71.9

3.2 実験結果

表 1 に示すように，実験結果からコード認識と音高推定を相互に組み合せることの有効性が示唆される．特徴量として NMF のアクティベーションから得られたクロマグラムを用いることで，通常のクロマグラムを用いるよりもコード認識率は 2.6 ポイント上昇した．これは，NMF によって倍音が除去され，より鮮明なピッチクラスの分布を表すクロマグラムが得られたためである．また，コード制約を与えた NMF によって得られたクロマグラムを用いた場合，コード認識率はさらに 0.3 ポイント上昇した．

提案手法によりコード認識率が期待ほど上昇しなかった理由の一つとして，The Beatles の楽曲では三和音のような比較的単純なコードを多用していることが挙げられる．単純なコードではクロマグラムは元からスパースであり，コードの事前分布の有無によって NMF の結果はほとんど変化しない．セブンスやナインスのような複雑なコードを多く含む楽曲に対しては，提案手法によるコード認識率の上昇が期待できる．

4. おわりに

本稿では，Bayesian HMM によるコード認識と Bayesian NMF による音高推定を組み合わせたコード認識手法について述べた．実験により提案手法の有効性が示されたが，コード認識率には更なる改善の余地がある．提案手法の効果を最大限に発揮するために，Bayesian NMF のハイパーパラメータを適切なものに調節する，もしくは学習する必要がある．一方で，提案手法によって音高推定の精度も向上することが期待される．したがって，今後はコード認識だけでなく，音高推定についても評価を行う予定である．

謝辞　本研究の一部は、科研費 24220006, 26700020, 24700168 および OngaCREST プロジェクトの支援を受けた．

参考文献

[1] C. Pérez-Sancho, D. Rizo and J. M. Inesta, “Genre Classification using Chords and Stochastic Language Models”, *Connection science*, Vol.21, No.2-3, pp.145–159, 2009.

[2] M. Goto, K. Yoshii, H. Fujihara, M. Mauch and T. Nakano, “Songle: A Web Service for Active Music Listening Improved by User Contributions”, *Proc. of ISMIR*, 2011, pp.311–316.

[3] T. Fujishima, “Realtime Chord Recognition of Musical Sound: A System using Common Lisp Music”, *Proc. of ICMC*, 1999, pp.464–467.

[4] K. Lee and M. Slaney, “Acoustic Chord Transcription and Key Extraction from Audio using Key-dependent HMMs Trained on Synthesized Audio”, *IEEE TASLP*, Vol.16, No.2, pp.291–301, 2008.

[5] M. Mauch and S. Dixon, “Approximate Note Transcription for the Improved Identification of Difficult Chords”, *Proc. of ISMIR*, 2010, pp.135–140.

[6] S. A. Raczyński, E. Vincent, F. Bimbot and S. Sagayama, “Multiple Pitch Transcription using DBN-based Musicological Models”, *Proc. of ISMIR*, 2010, pp.363–368.

[7] L. Oudre, Y. Grenier and C. Févotte, “Template-based Chord Recognition: Influence of the Chord Types”, *Proc. of ISMIR*, 2009, 153–158.

[8] O. Lartillot and P. Toiviainen, “A MATLAB Toolbox for Musical Feature Extraction from Audio”, *Proc. of DAFX*, 2007, pp.237–244.