Specmurt 分析と HMM を用いた音楽音響信号の調認識*

○齊藤翔一郎, 西本卓也, 嵯峨山茂樹 (東大情報理工)

1 はじめに

音楽信号から調 (tonality) を推定する技術は, 音楽情報検索, 自動採譜, 自動編曲, 自動伴奏などの多くの場面で重要な役割を果たす. 今回, 我々は音高情報取得の方法である specmurt 分析と, それにより得た特徴量をもとに和声をモデル化した HMM によって, 音響信号から調認識を行う手法を検討したので報告する.

2 調と和声進行のモデル化

2.1 動機

音楽の調性は, 長短 2 種の音階に基づく機能和声による音高の支配性を指す [1]. 音楽演奏の音響信号スペクトル時系列は, ある調性に支配された楽曲の中で遷移する和声の状態から出力される音響特徴量として観測されたものと見ることが出来る. そこで我々は, 調の隠れ変数である和声進行から生み出される音響特徴量をモデル化し, 音響特徴量が既知である時の逆問題として調の推定を行うことを考えた.

2.2 和声進行のモデル化

ここでは調性のある楽曲を扱うこととし, ある調性のもとに機能和声に基づく和声遷移により曲が構成されていると仮定する. 和声状態間の遷移は確率的と仮定し, 和声遷移をエルゴード的なマルコフモデルでモデル化する. 実際に観測されるのは音響特徴量であって和声の進行はその背後に隠れているため, これは隠れマルコフモデル (HMM) となる.

一般に和声学で扱われる和声は多種であり, その全てを和声状態として定義すると状態数が多くなりデータスパースネスの影響を受ける危険性がある. 調を認識するためには, 機能和声に基づく和声遷移を考慮すれば十分と考えられるが, T, S, D の機能のみでは現実の和声を反映しきれないとも考えられる. そこで, 今回は, 5 状態により和声進行をモデル化する. 今回この 5 状態は, 初期状態として I, II, IV, V, VI の 5 和音を与えて, 実際の楽曲データにより学習し, 最適化された和声状態および和声遷移を用いる.

このマルコフ性を定式化すると, ある調 K の楽曲で, 時刻と共に和声が $\mathbf{C} = \{c_1, c_2, \cdots, c_T\}$ と進行する確率 $P(\mathbf{C}|K)$ が

$$P(\mathbf{C}|K) = P(c_1|K) \prod_{t=2}^{T} P(c_t|c_{t-1}, K) \qquad (1)$$

と書けることに相当する.

2.3 調のモデル化

Fig.1 のように, マルコフモデルが長調と短調それぞれ 12 個ずつ 24 個あり, 入力である chroma vector 時系列はどれかひとつの調を通過して終端へと到達するようなモデルを考える. 楽曲が何調であるという問題は, 入力から何調のマルコフモデルを通って終端に達したかという問題として扱える. ここで調 K の用いられやすさを表す事前確率を $P(K)$ として考えることが出来るが, 今回は特別な事前知識を用いないことにする.

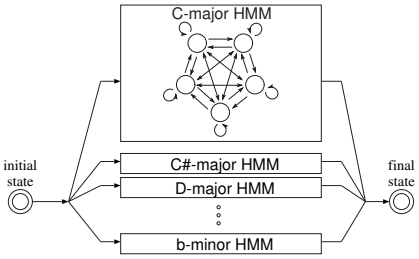
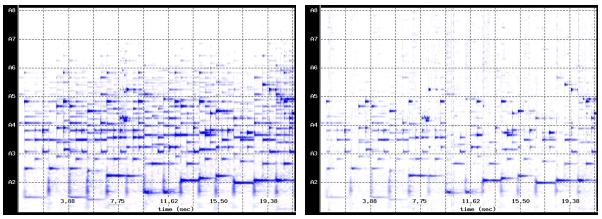


Fig. 1 調推定における調のネットワークモデル



(a) スペクトログラム (b)specmurt 分析結果

Fig. 2 specmurt 分析による倍音抑圧

3 音響特徴量としての Chroma Vector

3.1 Chroma vector の有用性

和声進行モデルから出力される音響特徴量は, 調の特徴を最大限反映したものであることが望ましい. 特徴量としてスペクトログラムを用いる場合, 次元数が多くなることや調に関係のない特徴までも含まれてしまう. またピッチ情報を用いる場合は和声に関する情報が失われてしまう.

音高情報のうち, オクターブ違いの同じ音階の成分を全て重ね合わせて 1 オクターブ内の半音階の 12 音の成分に縮約したものを chroma vector という. 楽曲の旋律や和声は, 全体をオクターブ単位で上下に平行移動しても調性は変化しないことから, オクターブ方向の分布の情報を取り除く chroma vector は, 調推定に必要な音高情報を圧縮していると言える.

3.2 Specmurt 分析による chroma vector

短時間解析から得た chroma vector には, 各楽音の成分以外に倍音成分が本来の和声の構成音と異なる成分として現れる. そこで和声を反映する音高情報を的確に取得するため, 倍音成分を除去することを考える. 我々は簡潔な計算で効率的に倍音成分を抑圧する手法である specmurt 分析 [2] を用いる. RWC 研究用音楽データベースのクラシック No.30(ノクターン No.2, ショパン) について specmurt 分析を行った結果を図 Fig.2 に示す.

我々は specmurt 分析結果を各半音階のオクターブ違いの強度を足し合わせ, 各要素の和を正規化することで得た chroma vector を音響特徴量として用いた. 今回は各音高の強度は各区間での強度の最大値を採用したが, 検討の余地があると言える.

3.3 Chroma vector の確率的変動

前節で求めた chroma vector は同一の調の同一の和声区間であっても同一値を持つわけではなく, 旋律

*Key Detection of Music Audio Signals via HMM through Specmurt Analysis. by SAITO, Shoichiro, NISHIMOTO, Takuya, SAGAYAMA, Shigeki (Graduate School of Information Science and Technology, The University of Tokyo)

中の非和声音や音強度バランスによって変動する．そこで, 各短時間周波数解析によって得られる chroma vector は, 確率な変動を伴って観測されると仮定する．

今回は, 正規化された chroma vector は多次元正規分布に従うと仮定する．即ち, 調 K で和声 c であるとき, chroma vector が $\mathbf{x}$ として観測される確率 $P(\mathbf{x}|c,K)$ は, 次の確率密度分布

$$\frac{1}{\sqrt{(2\pi)^{12}|\Sigma_{c,K}|}}\, e^{-\frac{1}{2}(\mathbf{x}-\mu_{c,K})^{\mathrm{T}}\Sigma_{c,K}^{-1}(\mathbf{x}-\mu_{c,K})} \qquad (2)$$

に従うとモデル化する．この分布のパラメータである平均 μ と分散行列 Σ は学習により推定することができる．

3.4 モデルの統合

調が K であり和声進行が C のように進行し, chroma vector の時系列データが $\mathbf{X}$ である確率を考える．各時刻の chroma vector が独立であるという仮定をおくと, 式 (1) とから

$$P(\mathbf{X}|\mathbf{C},K)\cdot P(\mathbf{C}|K)\cdot P(K) = \prod_{t=1}^{T}\Big\{P(\mathbf{x}_t|c_t,K)\cdot P(c_t|c_{t-1},K)\Big\}P(K) \qquad (3)$$

となる．ただし $P(c_1|c_0,K) = P(c_1|K)$ としている．前章で述べた HMM モデルにおいては, 式 (3) の $P(\mathbf{x}_t|c_t,K)$, $P(c_t|c_{t-1},K)$ はそれぞれ出力確率, 遷移確率に相当する．

4 HMM による調認識

4.1 MAP 推定による調認識

我々は, 観測された chroma vector の時系列 $\mathbf{X}$ に対して, 確率的に最も尤もらしい調 K を推定する．これは,

$$\hat{K} = \operatorname*{argmax}_{K} P(K|\mathbf{X}) \qquad (4)$$

ここで, Bayes の定理を用い, また観測からは隠れている和声にも注目すると, この事後確率は次のように書き換えられる．

$$\hat{K} \simeq \operatorname*{argmax}_{K,\mathbf{C}} P(\mathbf{X}|K,\mathbf{C})\cdot P(\mathbf{C}|K)\cdot P(K) \qquad (5)$$

この式の argmax の中身は式 (3) と等しくなり, 調推定は, 式 (3) を最大化する事後確率最大化問題に帰着される．具体的には, HMM のネットワーク上で時間同期 Viterbi 探索を行い, その最大尤度を図 1 の長調, 短調各 12 個の HMM で比較することで, 楽曲の調を推定できる．

4.2 HMM の学習

HMM の状態遷移確率と出力確率は学習データを用いて学習できる．24 種のすべての調ごとに学習すると, 学習データが不足し, かつ調に依存したモデル推定が行われて, すべての調を平等に扱えない恐れがある．そこで, 長調と単調のそれぞれをまず単一の調に移調してから学習する．ここでは長調の曲はハ長調, 短調の曲はイ短調に揃えている．そして学習されたハ長調とイ短調のパラメータを, chroma vector の空間上で回転して戻してやることで, 24 個の HMM のパラメータが更新される．ここで, 長/ 短調の各 12 の HMM の状態遷移確率や出力確率は調に依存しないと仮定し, 各パラメータは長/短調で値を共有させた．この更新されたパラメータによって, 調が未知のデータに対して探索を行い, 調の推定を行った．なお, 今回の評価実験では式 (2) の共分散行列は対角行列と仮定して出力確率を計算している．

Table 1　調認識の正解率 (MIREX, 数値:%)

acoustic feature	1-state HMM (Histogram)	5-state HMM		
		0.1*	0.05*	0.02*
wavelet	71.9	84.4	80.2	83.3
specmurt	76.0	**86.5**	**83.3**	**85.4**

Table 2　調認識の正解率 (RWC, 数値:%)

acoustic feature	1-state HMM (Histogram)	5-state HMM		
		0.1*	0.05*	0.02*
wavelet	89.1	**95.3**	**95.3**	**93.8**
specmurt	89.1	92.2	93.8	**93.8**

*: HMM 学習の σ^2 の初期値

5 実験

提案手法を実装し, 評価実験を行った．楽曲は 16kHz サンプリングのモノラル音声で, フレームシフトは 16ms, 周波数帯域 30-7609Hz で分析する．学習, 認識に用いた音楽音響信号は, MIREX2005[4] の “Audio and Symbolic Key Finding” のトピックで公開されている Training Data Set 96 曲と, RWC 研究用音楽データベースのクラシック音楽から転調していないと思われる 20 秒前後を切り出した曲 64 曲である．学習は open data で 4 分割の交差検証法を用いた．また, 学習の反復回数は 5 回である．各状態の初期値は I, II, IV, V, VI であるとし, chroma vector 出力確率の平均ベクトルは和声の構成音とそれ以外の音階が 5:1 となるように配分したうえで正規化したベクトルを用いている．

また, chroma vector 出力確率の共分散行列を対角共分散とし, 学習を行うときの初期値を $\sigma^2 = 0.1, 0.05, 0.02$ の 3 種類で行った．

Table 1 は MIREX, Table 2 は RWC のデータを用いた認識結果である．分散の初期値によって 3 つの結果を表示している．比較として, spectrogram を specmurt 分析をせずに用いた場合 (“wavelet”) と, HMM を用いずに曲全体を単一の chroma vector にして認識した場合 (“Histogram”) をあわせて掲載してある．

表に示すように, MIREX で 85%程度, RWC で 90% 以上の正解率を得ることができた．完全五度, 平行調, 同主調以外の誤りは全体の 1–2%に過ぎなかった．また両方の結果において, HMM の有効性が示されている．specmurt 分析に関しては, MIREX の結果では認識率が高くなっているが. RWC では手法によってどちらが優れているという有意な差は出ず, むしろ specmurt 分析を施さない方が認識率がよい傾向もある．これは RWC の方が複数の楽器を用いた楽曲が多いため, specmurt 分析の精度が低下しているということが原因のひとつとして考えられる．

6 まとめと展望

本稿では和声進行を HMM としてモデル化し, specmurt 分析を用いた chroma vector を音響特徴量として MAP 推定によって調を認識する手法を検討した．今後は特徴量やモデルを拡充させて, 転調の検出, 調認識の精度向上を目指したい．

参考文献

[1] 新音楽辞典 楽語, 音楽之友社, 1977.
[2] 亀岡他, “Specmurt 法による音楽信号の音高可視化における共通調波構造パターンの自動決定,” 音講論 (秋), 803–804, 2004.
[3] 川上, “HMM を用いた旋律への自動和声付けの研究,” 修士論文, 北陸先端科学技術大学院大学情報科学研究科, 2000.
[4] http://www.music-ir.org/mirexwiki/index.php/MIREX_2005