Vol.2010-MUS-86 No.13
2010/7/29

機能和声モデルによる音楽信号からの和声推定

上田 雄[†1] 小野順貴[†1] 嵯峨山茂樹[†1]

本研究では音楽信号からの和声推定を扱う。従来のモデルでは調を用いなかったり、転調を考慮しなかったりと、和声進行を十分に表現することが難しかった。そこで、従来の HMM による和声進行モデルを拡張し、機能和声に基づき転調を含む調と和音の認識を行うために、1) 隠れ状態として調と和音の組を持つ転調 HMM、2) 連続音声認識とのアナロジーから典型的な和声パターンを語彙として持つ和声語彙 HMM の 2 つのモデルを提案する。従来モデルとの比較実験によりその有効性を検証する。

Harmony estimation from music signals using functional harmony model

Yushi Ueda,[†1] Nobutaka Ono[†1]
and Shigeki Sagayama[†1]

In this report, we propose two models based on functional harmony, key-modulation HMM and harmony-vocabulary HMM, to estimate keys including key modulations and chords jointly from music signals. Evaluations of the models are made by comparing them with a conventional model.

1. はじめに

本研究では自動採譜や音楽情報検索への応用を目的として、音楽信号からの自動和声推定について扱う。西洋音楽などの調性音楽において、和声進行は重要な要素の一つであり、実演奏からの自動和声推定は自動採譜、カバー曲同定、音楽データベースの自動タグ付けなどの音楽情報検索 (MIR) の分野への応用への手掛かりとなる。たとえば、人間が音楽から採譜を行う際、和音を認識してからそれに合いそうな個々の音符を認識するというアプローチがある。これと同様のアプローチを計算機による自動採譜に適用することにより、自動採譜の精度を向上させることが考えられる。また、和音は楽曲の雰囲気を決定づける重要な要素であると考えられるため、和声進行を一つの指標として、カバー曲検索やユーザーの好みに合う楽曲の検索、楽曲の雰囲気のタグ付けなどに利用できるであろう。

我々が音楽を聴く際、楽曲の背後にある和声と調の関係を自然と感じられる。このことを理論づけたものを広い意味の機能和声理論と言うことができ、和声や調といった概念は機能和声理論に基づいている。ここで扱う音楽信号は和音の進行、調ともに未知であるが、我々は和音の進行を聴かなければ調を感じることはできず、逆に調を知らなければ和声学に基づく和音の進行を予測することはできない。したがって、これらは同時に推定することが必要だと考えられる。

和声推定問題に対し、連続音声認識との類似性から隠れマルコフモデル (HMM) の適用が主流となっている。HMM を適用した従来研究としては、まず川上らの研究[1] があげられる。川上らの研究では、与えられたメロディに対する自動和声付けという問題に対し、和声を隠れ状態としてそこから確率的にメロディが出力されるというモデル化がなされた。音楽音響信号を対象とした研究では、Sheh ら[2] はクロマベクトルと HMM を用い、以来クロマベクトルと HMM を用いた多くの関連研究がなされている。この研究では調は考慮せず、和音を隠れ状態としている (以降このモデルを「和音 HMM」と呼ぶ)。Lee らの研究[3] では調ごとに HMM を学習し、認識する際に全ての調で事後確率が最大となる和音系列を選択することにより、調と和音の両方の認識を可能にした。ただし、この研究では楽曲内での転調は考慮していない (以降このモデルを「調依存 HMM」と呼ぶ)。

本報告では機能和声に基づき、転調も含めた調と和音進行の両方を同時推定するための二つのモデルを提案する。一つは、隠れ状態を調と和音の組とすることで転調にも対応できるモデルで、本稿ではこのモデルを「転調 HMM」と呼ぶ。もう一つは楽曲中で頻出する和声進行を「語彙」として持つモデルで、以降「和声語彙 HMM」と呼ぶ。2 節ではまず音楽信号から特徴量を抽出する手法について述べる。3 節では従来モデルと提案モデルについて述べ、効果的なパラメータ共有法を紹介する。4 節では調と和音の認識実験により提案手法の評価を行う。5 節ではまとめと今後の展望を述べる。

†1 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

Vol.2010-MUS-86 No.13
2010/7/29

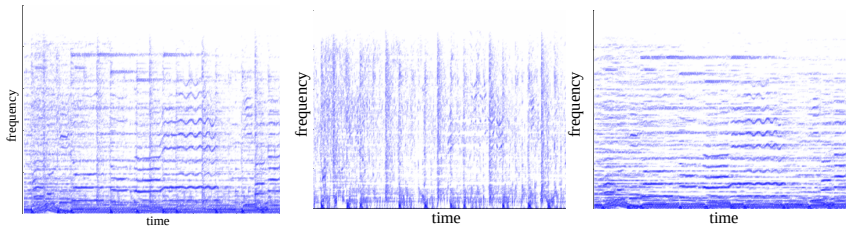

図 1 ポピュラー音楽の元のスペクトログラム $W(x,t)$(左)、調波音強調スペクトログラム $H(x,t)$(右)、打楽器音強調スペクトログラム $P(x,t)$(真ん中)

Fig. 1 The original spectrogram $W(x,t)$ (left), the harmonic-emphasized spectrogram $H(x,t)$ (right) and the percussive-emphasized spectrogram $P(x,t)$ (middle) of a popular music piece .

2. 特徴量抽出

2.1 調波音の強調

音楽音響信号では一般に打楽器音などの非調波な成分が含まれるが、これらは一定のピッチを持たないため、どの音が演奏されているかということが重要な和音認識において性能低下の原因となり得る。この問題に対し、**図 1** のように、信号のスペクトログラム $W(x,t)$ を調波成分 $H(x,t)$ と打楽器音成分 $P(x,t)$ に分離する宮本らによる手法[4),5)] を適用し、非調波音を抑圧することが考えられる[7)]。この手法では、調波音は時間方向に連結が強い成分であり、打楽器音は周波数方向に連結が強い成分であるというスペクトログラム上の滑らかさの異方性に着目をし、式 (1)、(2) の滑らかさのコストを定義し、式 (3) の目的関数 J を反復的に最小化することで分離を行っている。なお、m_H, m_P は W を調波成分・打楽器成分に分配する時間周波数マスクで、σ_P, σ_H は人手で実験的に定めるパラメータである。

$$\Omega_P = \frac{1}{2\sigma_P^2}\sum_{i,j}(\sqrt{P_{i-1,j}} - \sqrt{P_{i,j}})^2 \tag{1}$$

$$\Omega_H = \frac{1}{2\sigma_H^2}\sum_{i,j}(\sqrt{H_{i,j-1}} - \sqrt{H_{i,j}})^2 \tag{2}$$

$$\begin{aligned} J = & \sum_{i,j} m_P(x_i,t_j)W(x_i,t_j)\log\left(\frac{m_P(x_i,t_j)W(x_i,t_j)}{P(x_i,t_j)}\right) \\ & + \sum_{i,j} m_H(x_i,t_j)W(x_i,t_j)\log\left(\frac{m_H(x_i,t_j)W(x_i,t_j)}{H(x_i,t_j)}\right) \\ & - \sum_{i,j}(W(x_i,t_j) - P(x_i,t_j) - H(x_i,t_j)) + \Omega_P + \Omega_H \end{aligned} \tag{3}$$

2.2 クロマベクトル

和音は、さまざまなオクターブに渡って演奏されたり、いくつかの転回形や開離形、 密集形など様々な音高配置で演奏される。このような和音の音高配置によらない特徴量として、クロマベクトル[6)] がある。クロマベクトルは、式 (4) のようにパワースペクトルを半音ごとに複数オクターブ間で足し合わせることで得られる。ただし、$H(i,t)$ はスペクトログラムの周波数 bin i、時刻フレーム t でのパワー、I は取得するオクターブ数を表す。

$$p(k,t) = \sum_{i=0}^{I-1} H(12i+k,t) \tag{4}$$

スペクトログラムの取得に際して、STFT による時間周波数解析では低周波数で十分な周波数分解能を得るためには窓幅を広くとる必要があり、これにより、それほどの周波数分解能の必要のない高周波数の時間分解能まで下げてしまう。一方、定 Q フィルタバンクでは周波数と窓幅の比を一定に保つため、高周波数での時間分解能を落とすことなく低周波数での分解能を上げることができ、クロマベクトルを生成する際には定 Q フィルタバンクを用いて時間周波数解析を行う方が適していると考えられる。定 Q フィルタバンクの k 番目の中心周波数 f_k を平均律に従い

$$f_k = f_{min}2^{k/12} \tag{5}$$

とすることで、最低周波数 f_{min} からの半音毎の周波数 bin のスペクトログラムが得られる。

また、楽曲間の調律の相違に対処するためクロマベクトル候補を複数用意し、エネルギー最大である調律を選択することで調律の補正を行う[?)]。調律を補正したクロマベクトルは、式 (5) よりスペクトログラムの最低周波数 f_{min} を変化させることで求めることができる。f_{min} の候補は、基準とする周波数 f_0 を中心として $(f_0, f_0\cdot 2^{\pm 1/12n}, f_0\cdot 2^{\pm 2/12n}, \cdots)$ と、上下に $2^{1/12n}$ ずつ (つまり、$100/n$cent ずつ) 均等にずらした n 個とする。こうすること

で、あらゆるチューニングのずれに偏りなく対処することができると考えられる。

$$\hat{j} = \underset{j}{\operatorname{argmax}} \sum_{t=1}^{T} \sum_{k=1}^{12} p_j(k,t) \ , j = 1, \cdots, n \tag{6}$$

3. 機能和声モデル

3.1 HMM による和声進行のモデル化

特徴量系列 X が観測されたとき、その背後にある調系列 K、和音系列 C が求めるのが今回の問題である。この問題は事後確率最大化の観点から式 (7) と表現でき、さらにベイズの定理より式 (8) となる。

$$\{\hat{K}, \hat{C}\} = \underset{K,C}{\operatorname{argmax}}\, p(K, C|X) \tag{7}$$

$$\underset{K,C}{\operatorname{argmax}}\, p(K, C|X) = \underset{K,C}{\operatorname{argmax}}\, p(X|K, C)p(K, C) \tag{8}$$

ここで、観測特徴量の生成源として隠れマルコフモデル (HMM) を考える。隠れ状態は調と和声の組とし、各時刻で状態から特徴量が出力され、状態間で遷移するとモデル化する。和声進行には、和声学理論にあるように規則が存在すると考えられるため、現在の和声を推定する上で前までの和声を考慮することが必要であろう。そこで、ある時刻の和声は $n-1$ 時刻前までの和声に依存すると仮定し n-gram モデルにより表現する。ここでは簡単のため 2-gram モデルにより近似し、和声間の遷移確率を $p(k_t, c_t|k_{t-1}, ct-1)$ と表す。また、観測された特徴量と各和声の特徴量との近さの指標も必要となる。これは出力確率 $p(x_t|k_t, c_t)$ として表現する。

以上より、式 (8) は式 (9) と近似することができる。

$$\{\hat{K}, \hat{C}\} \simeq \underset{K,C}{\operatorname{argmax}}\, p(x_0|k_0, c_0)p(k_0, c_0) \prod_{t=1}^{T} p(x_t|k_t, c_t)p(k_t, c_t|k_{t-1}, c_{t-1}) \tag{9}$$

この最尤経路は Viterbi アルゴリズムにより効率的に求めることができる。

3.1.1 従来モデル

従来モデルの和音 HMM 、調依存 HMM の概念図を**図 2**、**図 3** に示す。和音 HMM は式 (9) において調を考慮しない場合と等価であり、和音間での遷移のみを考える。調を考慮しないことは和声進行のモデルとして粗い近似となっている。

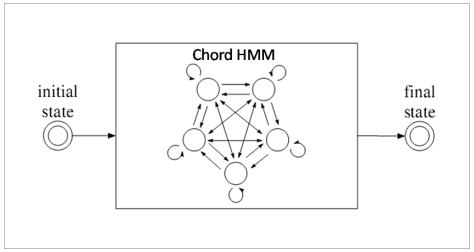

図 2 和音 HMM:和音間で遷移する
Fig. 2 Chord HMM: transitions between chords

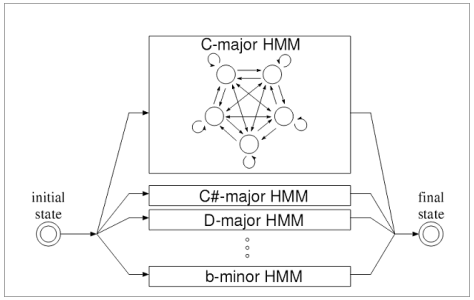

図 3 調依存 HMM: 同一調内の和音間で遷移する
Fig. 3 Key-dependent HMM: transitions between chords in the same key

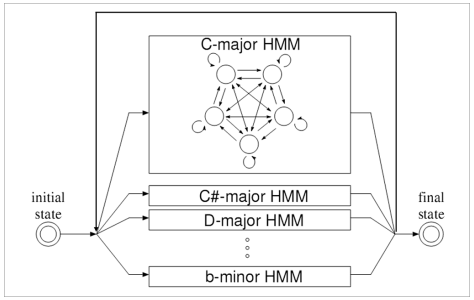

図 4 転調 HMM: 任意の調の和音間で遷移する
Fig. 4 Key-modulation HMM: transitions between chords in any key

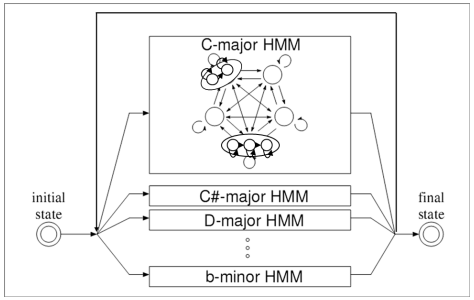

図 5 和声語彙 HMM: 任意の和声語彙間で遷移する
Fig. 5 Harmony-vocabulary HMM: transitions between any harmony vocabularies

調依存 HMM は式 (9) において調一定とした場合と等価である。このモデルでは調は考慮するものの、多くの楽曲に存在する転調を扱うことができない点でやはり現実に即しているとは言いがたい。

3.1.2 転調 HMM

概念図を**図 4** に示す。このモデルは式 (9) で表され、各時刻の調、和音を同時に推定することで転調も含めて推定することができる。楽曲を通しての最適解を求めるため、調と和音の相互依存性が考慮できていると言える。

3.1.3 和声語彙 HMM

和声進行にはカデンツのように、典型的なパターンが存在すると考えられる。そこで、音

声認識とのアナロジーから、これらのパターンを和声における語彙とみなすことで音楽的に妥当な認識結果を得られる可能性がある。このモデルを「和声語彙 HMM」と呼び、概念図を**図 5** に示す。しかし音声認識において語彙は既知であったのに対し、和声における語彙は必ずしも自明でない。そこで、データから学習することが考えられる。和声語彙の学習には様々な手法が考えられるが、本稿では以下のように行う。

- 2-gram 語彙
  もし $p(h_n|h_m)p(h_m) > p(h_n)p(h_m)$ ならば $h_m h_n$ を和声語彙に加える。
- 3-gram 語彙
  $h_l h_m$ が和声語彙であり、もし $p(h_n|h_m,h_l)p(h_m|h_l)p(h_l) > p(h_n|h_m)p(h_m|h_l)p(h_l)$ ならば $h_l h_m h_n$ を和声語彙に加える。
- 4-gram 語彙以降も同様に学習する。

ただし、h_n は n 番目の和声 $\{k_p, c_q\}$ に対応する。

3.2 モデルパラメータ共有

従来の和音モデルの状態数が (和音の種類数) であったのに対し、提案モデルでは (和音の種類数)×(調の種類数) だけ存在するため、学習に必要なデータが不足する可能性がある。そのため、和声の性質を利用したパラメータ共有を行うことにより、この問題に対処する。

3.2.1 音響モデル

和音の響きは調に依らず一定であると考えられるため、式 (9) は式 (10) と近似できる。

$$\{\hat{K},\hat{C}\} \simeq \underset{K,C}{\operatorname{argmax}}\, p(x_0|c_0)p(k_0,c_0)\prod_{t=1}^{T} p(x_t|c_t)p(k_t,c_t|k_{t-1},c_{t-1}), \tag{10}$$

更に、効果的なパラメータ共有を行うため各モード (長和音、短和音等) の響きはピッチシフトを除き一定であると仮定する。出力確率 $p(x|c)$ として単一正規分布を仮定すると (式 (11))、各モードの和音 N の平均 μ_N、分散 Σ_N に巡回シフト行列 S を掛けることにより各モードの全ての和音は同一のパラメータで表現することが出来る。尚、μ_N、Σ_N、S はそれぞれ式 (12)、式 (13)、式 (14) である。

$$p(x|c) = \frac{1}{\sqrt{(2\pi)^{12}|\Sigma_c|}}\exp\{-\frac{1}{2}(x-\mu_c)^T\Sigma^{-1}(x-\mu_c)\}, \tag{11}$$

$$\mu_N = S^N\mu_0, \tag{12}$$

$$\Sigma_N = S^N\Sigma_0(S^N)^T, \tag{13}$$

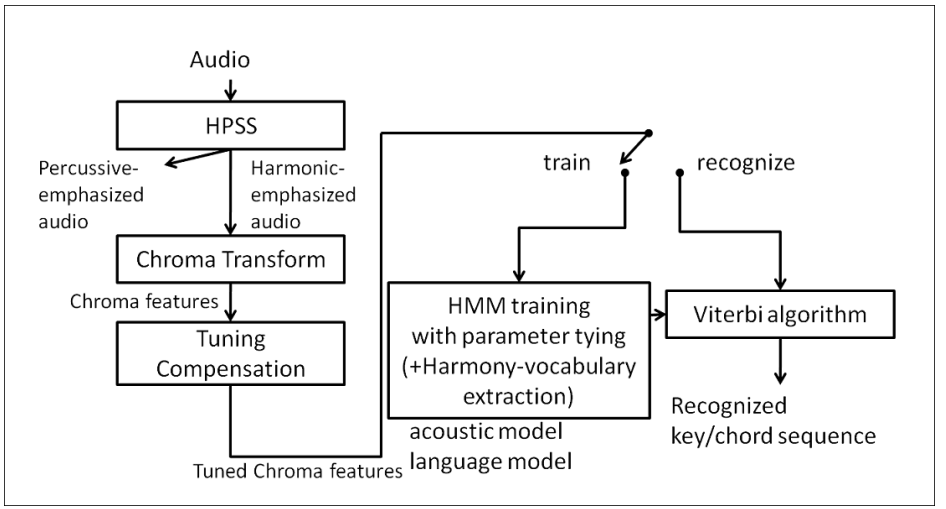

図 6 提案手法の概要
Fig. 6 Flow diagram of the proposed method

$$S = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \end{pmatrix}. \tag{14}$$

3.2.2 言語モデル

和音間の遷移は同一の和音遷移であっても異なる調では遷移確率は異なると考えられる。一方、機能和声に基づいて和音を調とその主音からの相対度数として捉えると、同一モードの調の和音記号間では遷移確率は等しくなるであろう。例えば C Major の V から I への遷移確率 $p(\mathrm{C:I|C:V})$ は G Major でも等しく、$p(\mathrm{G:I|G:V}) = p(\mathrm{C:I|C:V})$ とすることが出来る。異なる調間の和音遷移も考慮すると一般的に式 (15) のように書き表すことができる。

$$p(K_2,N_2|K_1,N_1) = p(K_2+M,N_2+M|K_1+M,N_1+M) \tag{15}$$

以上で議論した特徴量抽出、HMM の学習、認識の概要を**図 6** に示す。

Vol.2010-MUS-86 No.13
2010/7/29

表 1 オープンデータでの認識率
Table 1 Recognition results for open data

Model	Key Recog.	Chord Recog.
Chord HMM	–	79.9%
Key-modulation HMM	**75.8%**	**81.1%**
Harmony-vocabulary HMM	69.8%	80.8%

4. 評価実験

4.1 実験条件

和音 HMM、転調 HMM、和声語彙 HMM の調及び和音の認識性能を比較することにより、提案モデルの有効性を検証する。調依存 HMM との比較を行わないのは、転調 HMM において学習データに転調を含まない場合で学習したものと等価であるためである。

The Beatles の 12 枚のアルバム (“Please Please Me,” “With the Beatles,” “A Hard Day’s Night,” “Beatles for Sale,” “Help!,” “Rubber Soul,” “Revolver,” “Sgt. Pepper’s Lonely Hearts Club Band,” “Magical Mystery Tour,” “The Beatles,” “Abbey Road,” “Let It Be”) に含まれる 180 曲を用いて評価実験を行った。音楽音響信号は 11025Hz サンプリング、量子化ビット数 16bit、1 チャンネルであった。HMM の学習及び認識には Harte らによる和音ラベルの正解データ[9] を用いた。和音の種類は 12 の音名それぞれにおける major/minor の 24 種類に無和音 (無音や発話に対応) を加えた 25 種類であった。調の種類は 12 の音名それぞれにおける major/minor の 24 種類であった。認識率は調と和音それぞれに対し全 180 曲での (正解フレーム数)/(全フレーム数) で計算した。この計算の際には正解データ中での無和音の区間は除外された。性能評価では二つの実験が行われた。

(1) オープンデータでの実験
学習・認識は 3-fold cross-validation により行い、8 枚のアルバムで HMM を学習し、残り 4 枚のアルバムの認識を行い、それを 3 回繰り返すことにより、全 180 曲の調と和音認識結果を得た。

(2) クローズドデータでの実験
学習・認識は全 180 曲で行われ、全 180 曲の調と和音認識結果を得た。

4.2 実験結果: オープンデータでの実験

実験結果を**表 1** に示す。提案モデルはどちらも従来の和音 HMM の和音認識率を上回ることを確認した。これにより、調を考慮することの有効性が示された。中でも、調、和音両方の認識率において転調 HMM の認識率が最も高かった。各モデルの各楽曲でのヒストグラムを**図 7** に示す。ここから、提案モデルがわずかに認識率 80 %以上の楽曲が増加している傾向が分かる。調認識の傾向として、多くの転調個所を検出できていた一方で、転調しない個所を転調していると誤る個所も存在した。**図 8** に各モデルでの認識例を示す。

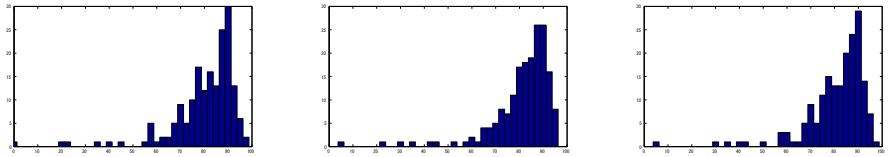

図 7 各モデルによる、各曲の認識率ヒストグラム (縦軸：楽曲数、横軸：認識率): 和音 HMM (左)、転調 HMM (真ん中)、和声語彙 HMM (真ん中)
Fig. 7 Recognition rate histogram of each model (vertical axis: # of songs, horizontal axis: recognition rate): Chord HMM (left), Key-modulation HMM (middle), Harmony-vocabulary HMM (right).

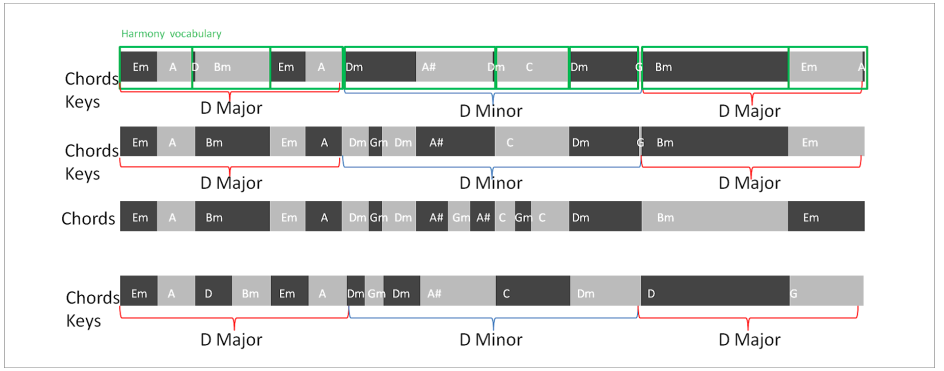

図 8 “The fool on the hill” の各モデルでの認識例: 和声語彙 HMM (第 1 列)、転調 HMM (第 2 列)、和音 HMM (第 3 列)、正解ラベル (第 4 列)
Fig. 8 Recognition results for “The fool on the hill”: Harmony-vocabulary HMM (1st row), Key-modulation HMM (2nd row), Chord HMM (3rd row), Reference label(4th row)

4.3 実験結果: クローズドデータでの実験

和声語彙 HMM がオープンデータで高い認識性能を得られなかったのは、モデルの複雑さによる学習データ不足のためだと考えられる。そこで、クローズドデータによる実験を

表 2 クローズドデータでの認識率
Table 2 Recognition results for closed data

Model	Key Recog.	Chord Recog.
Chord HMM	–	79.9%
Key-modulation HMM	84.4%	82.1%
Harmony-vocabulary HMM	**87.0%**	**82.9%**

行った。

実験結果を**表 2** に示す。調、和音両方の認識率において和声語彙 HMM の認識率が最も高かった。これにより、十分な学習データが存在した場合の和声語彙 HMM が有効である可能性が示唆された。

5. おわりに

本報告では機能和声に基づく 2 つのモデルを提案した。隠れ状態を調と和音の組とする転調 HMM、典型的な和声パターンを語彙として持つ和声語彙 HMM により転調を含む楽曲の認識が可能となった。また、学習データ不足に対処するためのパラメータ共有を紹介した。従来モデルとの比較実験によりその有効性を確認した。

今後は和声語彙の抽出や和声語彙 HMM の言語モデルの学習法について検討していきたい。また、RWC 音楽データベースのクラシック音楽に対する和声ラベル[10] が存在するため、それを用いた実験を行う予定である。

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) 川上隆他, “隠れマルコフモデルを用いた旋律への和声付け,” 平成 11 年電気関係学会北陸支部大会講演論文集, F-61, p. 361, 1999.
2) A. Sheh *et al.*, “Chord segmentation and recognition using EM-trained hidden markov models,” *Proc. ISMIR*, pp. 183–189, 2003.
3) K. Lee and M. Slaney, “Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio,” *IEEE Trans. ASLP*, vol. 16, no. 2, pp.291–301, 2008.
4) 宮本賢一他, “スペクトログラムの滑らかさの異方性に基づいた調波音・打楽器音の分離,” 日本音響学会春季研究発表会講演論文集, pp. 903–904, 2008.
5) N. Ono *et al.*, “Harmonic and Percussive Sound Separation and its Application to MIR-related Tasks,” Advances in Music Information Retrieval, ser. Studies in Computational Intelligence, Z. W. Ras and A. Wieczorkowska, Eds. Springer, 274, pp.213-236, Feb., 2010.
6) T. Fujishima, “Real-time chord recognition of musical sound: A system using common lisp music,” *Proc. ICMC*, pp. 464–467, 1999.
7) 内山裕貴他, “調波音/打楽器音分離手法を用いた音楽音響信号からの自動和音認識,” 情報処理学会研究報告, 2008-MUS-76, pp. 137–142, 2008.
8) 上田雄他, “調波音／打楽器音分離手法とチューニング補正手法を用いた音楽音響信号からの自動和音認識,” 情報処理学会研究報告, 2009-MUS-81, 2009.
9) C. Harte *et al.*, “Symbolic representation of musical chords: A proposed syntax for text annotaions,” *Proc. ISMIR*, pp. 66–71, 2005.
10) 川上大輔他, “和声ラベルデータの作成と和声進行の統計解析,” 情報処理学会研究報告, 2010-MUS-84, 2010.