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

内山裕貴[†1] 宮本賢一[†1] 西本卓也[†1]
小野順貴[†1] 嵯峨山茂樹[†1]

本稿では, 音楽情報検索や自動採譜等への応用を目的とした, 音楽音響信号から和声進行を認識する問題を扱う. ポピュラー音楽等の音楽音響信号は, 一般に打楽器等の非調波的な音を含み, 音楽の調波的な要素である和音の認識が難しい. そのため, 我々は, 本研究室で開発した調波音・打楽器音分離手法を用い, 調波音を強調したクロマベクトルを特徴量に用いる手法を提案した. この手法を, 多くの楽曲を用いて評価を行い認識率が向上することを確認した.

Automatic Chord Detection Using Harmonic/Percussive Sound Separation from Music Acoustic Signals

Yuki Uchiyama,[†1] Kenichi Miyamoto,[†1] Takuya Nishimoto,[†1]
Nobutaka Ono[†1] and Shigeki Sagayama [†1]

This paper describes the automatic chord detection for music information retreval or automatic music transcription. It is hard to detect chord progression from music acoustic signal such as popular music because they include non-harmonic sounds for example drum sounds. So we proposed a method using chroma emphasized harmonic sounds generated by harmonic/non-harmonic sounds separation developed our laboratory. In this papaer we evaluate this method using many song.

1. はじめに

本稿では, 音楽情報検索や自動採譜等への応用を目的とした, 音楽音響信号から和音を認識する問題を扱う. 和音は西洋音楽のような調性音楽においては音楽の構造を決める基本的な要素であり, 音楽情報処理や音楽信号処理の分野で様々な応用が考えられる. 音楽情報処理では, 音楽情報検索 (MIR) や音楽分類等, 音楽の内容に基づいた解析に対し和音は有用である. 具体的には, 編曲やアレンジされた音楽がどの音楽から派生したのかという音楽のカバーソングの識別問題は音楽の検索や著作権保持などの利用が考えられるが, カバーソングと原曲は和音が変化しないことが多く, 和音が重要な手掛かりとなる. また音楽信号処理では, 自動採譜の前処理としての利用が考えられる. 人間が採譜をする場合に先に和音を求めてから, 一つ一つの音を聞いていく場合があり, 自動採譜においても同様の利用できる可能性がある.

和音認識の従来研究は, シンボリックな旋律に対し和声モデルに隠れマルコフモデル (HMM) を用い, 和声付けを行った川上らの研究[5] がある. ポピュラー音楽の音楽音響信号に対して和音認識をした研究は Sheh らの研究[1] がある. Sheh らは, クロマベクトル[8] と呼ばれる, 12 半音毎のパワーをオクターブ間で足し合わせたベクトル時系列を特徴量に用い, 和声モデルに HMM を用いて和音認識を行った. また, 一般的に手に入りやすい和声進行と音楽音響信号から Baum-Welch アルゴリズムを用いることにより, 音楽音響信号と和声進行のアライメントを人手でつけることなく, 学習した. その他にも, クロマベクトルと HMM に加えて, 音楽知識の導入[4] や, クロマベクトルに対する倍音の影響のモデル化[3] などがある.

本稿で提案する手法もクロマベクトルと HMM を用いるが, 我々は, 音楽音響信号を扱う上で困難な点の一つに, 音楽は一般的に打楽器音やアタック音のようにピッチを持たない非調波的な音を含むことが多いという点に着目した. 和音は音楽の調波的な要素であり, 非調波的な音によりの認識は困難になると考えられる. そこで我々は, 宮本らの調波音・打楽器音分離手法[7] により得られた調波音から求めたクロマベクトルを特徴量に用いる手法を提案した[9]. 本稿ではその手法により多くの楽曲を用いて評価実験を行い, 認識率が向上することを確認したことを述べる.

本稿の構成は, 以下のとおりである. 第 2 章では, 調波音・打楽器音分離手法を用いたクロマベクトルの特

†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

徴量と HMM を用いた, 和音認識アルゴリズムについて述べる. 第 3 章では, この手法の効果を確かめる実験を行い, 評価を行う. 第 4 章でまとめと今後の展望を述べる.

2. 調波音を強調したクロマベクトルによる和音認識

2.1 調波音を強調したクロマベクトル

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[8] がある. クロマベクトルは, パワースペクトルを半音名ごとに複数オクターブ間で足し合わせることで得られ, それぞれの半音名のパワーを表している.

さて, 本稿で主に対象とするポピュラー音楽は, 一般に打楽器音やアタック音のような非調波的な音を含むが, これらの音はピッチを持たない音であり, 音楽の調波的な要素である和音を認識する際に妨げとなる. そこで我々は, 調波音を強調し打楽器音を抑制するために宮本ら[7] の調波音・打楽器音の分離を用いた. この手法は, 「スペクトログラムにおいて, 打楽器は時間方向に急峻に変化するが周波数方向には滑らかであり, 調波音は逆に周波数方向に急峻であるが時間方向には滑らかである」という点に着目し, スペクトログラム上で, 時間方向と周波数方向の滑らかさを用いて, 調波音と打楽器音を分離する. この手法では入力信号のスペクトログラム $W(x,t)$ を調波音のスペクトログラム $H(x,t)$ と非調波音のスペクトログラム $P(x,t)$ に分離する. このとき滑らかさの制約として

$$\Omega_{\mathrm{H}} = \frac{1}{2\sigma_{\mathrm{H}}^2} \sum_{i=1}^{I} \sum_{j=1}^{J-1} \left(\sqrt{H(x_i, t_{j+1})} - \sqrt{H(x_i, t_j)} \right)^2$$

$$\Omega_{\mathrm{P}} = \frac{1}{2\sigma_{\mathrm{P}}^2} \sum_{i=1}^{I-1} \sum_{j=1}^{J} \left(\sqrt{P(x_{i+1}, t_j)} - \sqrt{P(x_i, t_j)} \right)^2$$

を導入している. この制約の和と, 分離されたエネルギー分布と, $H(x,t)$ 及び $P(x,t)$ との距離の和を目的関数として, 反復推定により調波音と打楽器音を分離できる. ここで, Ω_{H} と Ω_{P} はあらかじめ与えるパラメータであり, スペクトログラムのエネルギーが調波音に分類されやすいか, 打楽器音に分類されやすいか, に関わる.

特徴量の計算は以下のように行う. まず入力信号から短時間フーリエ変換によりスペクトログラム $W(x,t)$ を得て, 調波音・打楽器音分離手法を用いて調波音のスペクトログラム $H(x,t)$ を計算する, そこから得られた調波音の信号を音量で正規化して定 Q フィルタバンクによりスペクトログラム $H'(x,t)$ を計算する. x は周波数軸上に対数でとり, 各 x に対応する周波数 $f(x)$ は

$$f(x) = f_{\mathrm{ref}} \cdot 2^{\frac{(x - x_{\mathrm{min}})}{12}} \mathrm{Hz}$$

とした. f_{ref} はチューニングによる基本周波数で, x_{min} は x のオフセットである. ここで, 定 Q フィルタバンクを用いて時間周波数解析を行うのは, クロマベクトルを計算するために, 周波数方向に対数のピッチに合わせて解析することが容易であること, 高音での時間分解能を下げることなく, 低音の周波数分解能をあげることができるからである. 低音で演奏されるベースなどは和声内音を演奏することが多く, これらの音域まで含むことは和音認識において有効である. 最後に $H'(x,t)$ を周波数方向にオクターブ間で足し合わせることにより

$$p(k,t) = \log \left(\sum_{i=0}^{I-1} H'(12i + k, t) \right)$$

とクロマベクトル $p(k,t)$ を計算する. I はクロマベクトルを計算するオクターブ数である. クロマベクトルの各次元が等価になるように, 整数倍のオクターブで計算した.

2.2 和音の HMM によるモデル化

和音は, 和声内音がそのまま演奏されるわけではなく, 和声内音の省略や非和声音の挿入があるため, フレームごとにみて認識するのは難しい. そこで, 現在の和音を認識するのに周辺の時刻の和音を手がかりとして利用することを考える. 和音は時間ごとにランダムに出現するのではなく同じ和音がある程度の時間連続したり, 和音間での遷移のしやすさには偏りがあったりするという性質を利用する. このように和音の性質を用いた定式化は川上ら[5] によって行われた.

和音を, 音楽に直接表れないが隠れて存在する状態と捉え, 和音のモデルに関して以下の仮定をする.

- 和声進行は, 現在の和音が $N-1$ 個前までの和音に依存する N-gram による確率過程である.
- クロマベクトルは, その時刻の和音名によって確率的に生成される.

1 番目の仮定については, 遠く離れた和音の影響をあまり受けないことを考えると近似として妥当である. 特にひとつ前の和音による影響が支配的であると考え, 2-gram によってモデル化する. なお, ここでは調を考えずに和音を考えているが, 和音間の遷移のしやすさは調によって異なる. そのため和音間の遷移は調を考えた場合に比べ, 粗い近似のモデルとなるが転調のある曲にも容易に対応できるという利点があり. このモデルを用いた. 調を考慮したモデルに関しては, 検討中である.

2 番目の仮定は, 和音が決まった時に和声内音は表れやすく, 非和声音が表れにくいことを表す事ができる. 実際の楽曲では旋律等の音の動きにはある偏りがあったり, 和音の開始時刻と終了時刻では演奏されやすい音が異なる事も考えられる. 特に後者については

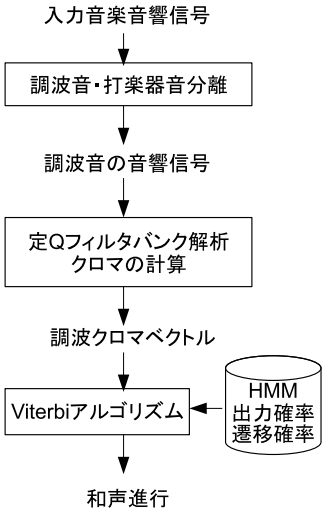

図 1 和音認識アルゴリズム

今後検討予定である.

以上のような仮定を置いて, 和音のエルゴディックな HMM を用いて定式化する. 入力の調波音によるクロマベクトル系列を x, 求める和声進行系列を c とおくと, 和音認識問題は, ベイズの定理により

$$\operatorname*{argmax}_c p(c|x) = \operatorname*{argmax}_c p(x|c)p(c)$$

となる. ここで, 和音の遷移について, HMM で近似することにより,

$$\operatorname*{argmax}_c p(c|x)$$

$$= \operatorname*{argmax}_c p(x_0|c_0)p(c_0)\prod_{t=1}^{T} p(x_t|c_t)p(c_t|c_{t-1})$$

となり, 和音毎のクロマベクトルの出力確率 $p(x_t|c_t)$ と, 和音間の遷移確率 $p(c_t|c_{t-1})$ をあらかじめ持っていれば, 最尤となる和声進行系列 c を計算できる. この最尤経路は Viterbi アルゴリズムにより求めることができる.

それぞれの和音のクロマベクトルの出力確率 $p(x_t|c_t)$ として, 我々は単一の正規分布を仮定した. これは, 対数クロマベクトルが実験的に正規分布に近い分布になることが多かったこと, 計算がしやすいことが理由である. また, クロマベクトルの各次元は独立ではないため, 非対角要素も非零である. これは, ある和音が演奏されているときには, その和声内音のように同時に演奏されやすい音の組合せがあること, また倍音の影響によりある音が演奏されているときに別のピッチでもパワーが存在するためである.

本章で述べたアルゴリズム全体を図 1 にまとめた.

3. 評価実験

3.1 実験条件

前章の調波音・打楽器音分離手法を用いた和音認識アルゴリズムを評価するための実験を行った. 学習及び認識には 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 曲の楽曲を用いた. これらの楽曲を用いたのは, 先行研究[3] 等で用いられており, 今後性能比較することができるためである.

楽曲はすべてヴォーカルを含む多重音曲であり, 打楽器有り・無しの曲, 転調のある曲を含む. CD のデータをモノラル化し, 11025Hz にダウンサンプリングしたものを入力の音楽音響信号とした. 今回の実験は, 調波音・打楽器音分離を用いたクロマベクトルの効果を確認するための実験であるから, 扱う和音は長三和音, 短三和音の 24 種類とした. また, 対象曲の先頭, 末尾, 曲中には無音やセリフといった和音のない区間があり, 1 状態のコード無し区間としてモデル化した. なお, 正解率の計算のときには, この無音区間は和音ではないためカウントしていない.

実験は次の 2 つの条件で行った.

- ドラム有り・無しの曲を含む 180 曲
- ドラム有りの曲 166 曲

対象の 180 曲には打楽器なしの曲があり, 調波音・打楽器音分離の効果を確認するために, ドラム有りの曲のみの実験も行った.

学習・正解ラベルは C. Harte らの作成した[2] ものを利用した. このラベルデータは曲全体の和声進行だけでなく, 各時刻でどの和音であるかというアライメント情報も含むため, HMM の出力確率と遷移確率の学習は, 状態系列が与えられている場合のパラメータの最尤推定を行った. 状態系列が既知のため, 精度よくパラメータ学習ができる. また, このラベルは, 長三和音, 短三和音以外の和音も含むため, それ以外の和音については, 図 1 のように, 根音はラベルの根音を用い, 第三音が短三度ならば短三和音, それ以外は長三和音とみなした.

それぞれの条件に対し, データセットを 2 つに分け, ある 90 曲 (84 曲) を用いて学習を行い, 残りの 90 曲 (84 曲) を認識し, 学習データと認識データを入れ替えて, 同様に実験を行った.

また, 調波音・打楽器音分離のパラメータ Ω_{H}, Ω_{P} は, 実験的に設定した. 良い性能を示したパラメータでは, 非調波音に分類される音が多く聞こえ, 打楽器やアタック音の多くだけでなく細かい動きのメロディーも非調波音に分離され, 調波音はベースや比較的長い

表 1 和音のクラスタリング

和音の例	みなす和音
C maj, C sus4, C aug, etc	C maj
C min, C dim, C min7, etc	C min

表 2 和音認識の正解率

楽曲数	混合信号	調波信号
180 曲	198017 フレーム (70.7%)	211983 フレーム (75.7%)
166 曲	183704 フレーム (71.2%)	195729 フレーム (75.9%)

音を演奏するギター音がよく聞こえた. なお , このパラメータを用いて分離された非調波音/調波音のエネルギー比率は全曲平均で 1.09 となった.

時間周波数解析のパラメータ $f_{\mathrm{ref}} = 440.0$Hz とし, 解析する音域を A1 (55.0Hz) 〜G#6 (3322.4Hz) とするために, $x_{\min} = 36$, $I = 6$ とした. この音域では, 対象とする楽曲のほとんどの基本周波数を含む.

認識時の文法モデル/音響モデルの重みは実験的に 4 とした.

3.2 結果と考察

評価実験の結果を表 2 に示す. 認識率はいずれの条件でも全曲で平均すると上昇している. また, 楽曲数を 180 曲としたときの, 混合信号と調波信号で各曲の認識率の度数分布は図 3 のようになった.

混合信号に比べ, 調波信号を用いた時には 148 曲で認識率が向上し, 31 曲では認識率が低下した. 調波信号を用いることで, 認識率が高い曲が増えている. 混合信号・調波信号どちらの場合も極端に認識率が低い曲があるが, これらの曲は西洋音楽や調性音楽ではないためである.

楽曲数が 166 曲の場合も, 調波信号を用いることで認識率が向上している. この原因としては, ドラムを含まない曲でも楽器のアタック音があり調波音・打楽器音分離によってこれが抑制されたことや, 調波音・打楽器音分離手法が打楽器音だけでなく, 動きの細かいメロディーも抑制し, 結果的にベース音やコードを演奏している楽器が強く残っていることが原因として考えられる.

学習した出力確率分布は図 2 のようになり, 混合信号に比べて調波信号では和声内音の平均値が高くなり, 分散共分散の値も全体的に小さくなっている. これにより, 和声内音と非和声音がより明確に区別できるようになったと考えられる. なお, C の和音であるが, 和声内音の (C 音, E 音, G 音) 以外に, B 音や D 音が大きな平均を持つが, これは C の和音の区間に B 音や D 音が演奏されやすい, E 音や G 音の倍音としてパワーを持つ, 和音 CM7 (C 音, E 音, G 音, B 音) 等の区間を和音 C とみなしていることなどが原因として考えられる.

認識結果が上がった曲と下がった曲の例ついて認識結果を図 4 に示す.

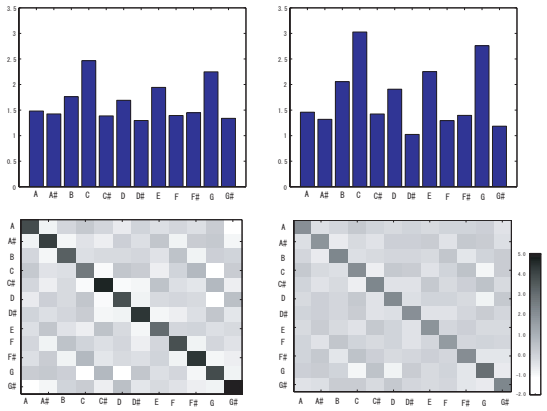

図 2 出力確率分布

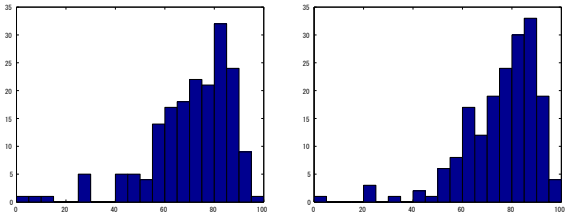

図 3 曲別の認識率

図 4 の (a) , (b) は認識率の向上が大きかった “Anna (Go To Him) ” の認識結果であり, この曲は認識率が 25.9%から 57.8%に向上した. 混合信号では認識できなかった D と Bm の区間がほぼ正しく調波信号では正しく認識できている. クロマグラムをみると, 混合信号では各音に渡ってパワーが存在するが, 調波信号では和声内音のパワーが強く, 認識率が向上していると考えられる.

図 4 の (c) , (d) は認識率の低下が大きかった “I Wanna Be Your Man” の認識結果であり, この曲は認識率が 84.9%から 32.5%に低下した. この曲は全体を通して E のコードが多く出て来るが調波信号ではその多くを Em と認識してしまい, 認識率が低下した. クロマグラムを見ると, E の区間で第三音の G#音 のパワーが弱く, 明確に演奏されていないため認識が困難である. 今回の学習セットの場合, Em の和音の区間では, 非和声音の D 音のパワーの平均も大きく観測された. クロマグラムでも D 音にパワーがあるのが分かり, 調波信号ではこれにつられて, Em と認識したと考えられる. 混合信号でも D 音にパワーがあるが, 混合信号では出力確率の分散が大きいため,D 音の影響をあまり受けずに, 文法の影響により Em よりは E と認識したことが考えらる. 実際, 混合信号で文法モデル/音響モデル比率を下げていくと E の和音の区間

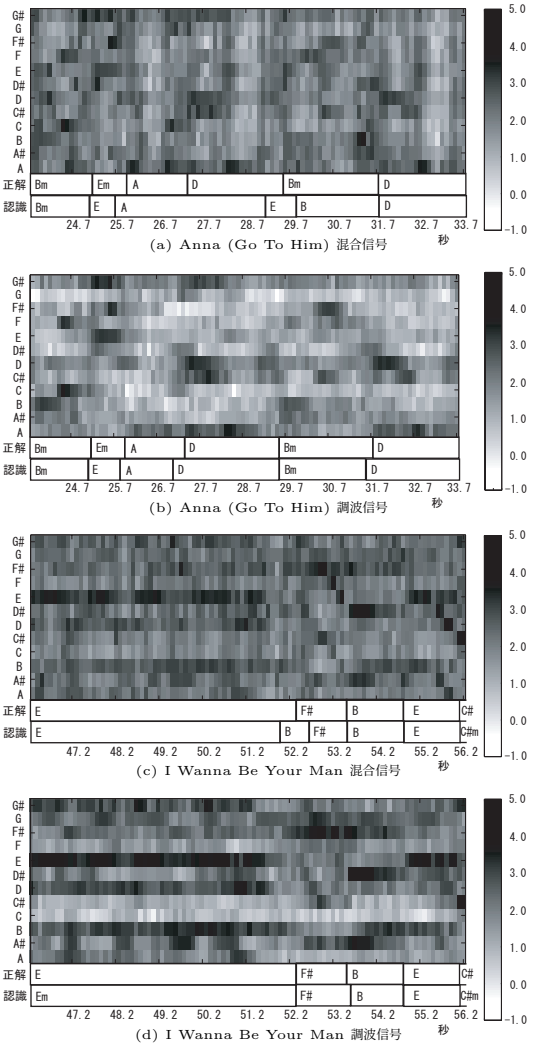

図 4 認識率結果詳細

が Em の和音と誤認識されることがあった.

また, このアルゴリズムが The Beatles 以外の曲に対しても有効であることを示すために, RWC 研究用音楽データベース[6] の楽曲に対しても, 同様に実験を行ったところ, その認識率は表 3 のようになった. この 7 曲はすべてヴォーカルを含む多重音楽曲で RWC-MDB-P-2001 No.74 以外はドラムを含む. 学習データに関しては The Beatles の楽曲 90 曲を用いた. 7 曲すべてで認識率が向上した. 全体的な認識率が, 表 2 と比べて低いのは学習データが The Beatles の楽曲であり, The Beatles らしい和声進行やクロマベクトルを学習しているためだと考えられる. これについては今後検討予定である.

表 3 和音認識の正解率 (RWC データベース)

楽曲	混合信号	調波信号
RWC-MDB-P-2001 No.14	58.5%	71.4%
RWC-MDB-P-2001 No.17	42.3%	70.3%
RWC-MDB-P-2001 No.40	45.2%	50.0%
RWC-MDB-P-2001 No.44	47.3%	67.2%
RWC-MDB-P-2001 No.45	64.3%	68.5%
RWC-MDB-P-2001 No.46	65.7%	75.2%
RWC-MDB-P-2001 No.74	74.7%	75.9%
合計	56.0%	68.1%

4. おわりに

4.1 まとめ

本稿では, 音楽音響信号から自動和音認識の性能向上のために, 特徴量に調波音を強調したクロマベクトルを用いる手法を提案し, 評価実験により性能が向上することを示した. 認識率を向上させるためには, 次節で示す展望が考えられる.

4.2 今後の展望

4.2.1 打楽器音の利用

本稿では調波音・打楽器音分離手法によって得られた調波音を用いる手法を述べたが, 今後の展望として打楽器音の利用について検討する. 音楽音響信号からの和音認識では, 和音名称に加えて和音境界も未知であり, 前章のシステムでは和音境界誤りが和音認識誤りの原因の一つとなっていた. この問題を解決するために, 和音境界は拍の位置に置かれて打楽器音が重畳する確率が高いという仮説のもとに, 打楽器音を和音境界の手がかりとして利用することを考えたい.

調波音・打楽器音分離によって得られた打楽器音のパワーを特徴量の一つに用いることを考える. 実際には拍の上でも打楽器が演奏されなかったり, 拍のタイミング以外でも打楽器が演奏されたりするので, 調波音や前後の関係を確率的に考慮する必要があり, 図 5 のような HMM が利用できる. HMM の状態として (C, major, 拍有り) , (C, major, 拍無し) , ··· のように和音 24 種と拍有り・無しによる合計 48 状態を考える. 同一和音の拍有りの状態と拍無しの状態は left-right 型の HMM をなし, 和音間ではどの和音にも遷移しうるエルゴディックな HMM とする. 状態出力は調波クロマベクトルと打楽器特徴量の 13 次元とし, 拍有りの状態では打楽器特徴量が大きく, 拍無しの状態では打楽器特徴量が小さいような分布を用いることにより, 打楽器が強い場合には拍の位置である可能性が高く, 和音が変わりやすいという現象が扱える. 今後, 実験検証する予定である.

4.2.2 調の利用

今回の定式化では HMM の状態を 24 種類の和音とし, 和音の遷移は 1 状態のマルコフ過程と近似したため, 和音間の遷移確率は二つの和音だけで決まること

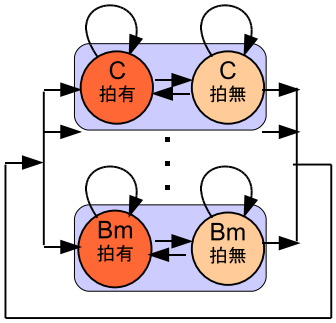

図 5 打楽器音を考慮した HMM

になる．我々の対象とする調性音楽では調があり，同一和音間でも調によって遷移のしやすさは異なると考えられる．調が未知であり転調のある曲を含むため，このようなモデル化を行ったが，調を用いることによりより実際の楽曲即したモデル化が考えれ，今後の課題である．

謝辞 本研究の一部は，科学技術振興機構 CREST の補助を受けて行われた．

参 考 文 献

1) A. Sheh *et al.*, “Chord segmentation and recognition using EM-trained hidden markov models,” Proc. ISMIR, pp. 183–189, 2003.
2) C. Harte *et al.*, “Symbolic representation of musical chords: A proposed syntax for text annotaions,” Proc. ISMIR, pp. 66–71, 2005.
3) H. Papadopoulos *et al.*, “Large-scale study of chord estimation algorithms based on chroma representation and HMM,” Proc. CBMI, pp. 53–60, 2007.
4) J. P. Bello *et al.*, “A robust mid-level representation for harmonic content in music signal,” Proc. ISMIR, pp. 304–311, 2005.
5) 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.
6) M. Goto *et al.*, “RWC music database: Popular, classical, and jazz music databases,” Proc. ISMIR, pp. 287–288, 2002.
7) 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, 2008.
8) T. Fujishima, “Real-time chord recognition of musical sound: A system using common lisp music,” Proc. ICMC, pp. 464–467, 1999.
9) 内山裕貴他, “調波音を強調したクロマに基づく音楽音響信号からの自動和音認識,” 日本音響学会春季研究発表会講演集, pp. 901–902, 2008.