Vol.2015-MUS-108 No.1
2015/8/31

音楽音響信号に対する歌声・伴奏音・打楽器音分離に基づくコード認識

丸尾 智志[1,a)]　池宮 由楽[1,b)]　糸山 克寿[1,c)]　吉井 和佳[1,d)]

概要：本稿では，音楽音響信号に対する歌声・伴奏音・打楽器音分離に基づくコード認識手法を提案する．従来法では，打楽器音がコード認識に与える悪影響を取り除くために，音楽音響信号を調波音のみに分離してコード認識を行っている．この方法の問題点は，調波音に含まれる歌声の影響を考慮していないことである．歌声はコードとは無関係の音を演奏することも多い上に，伴奏に比べると音量が大きいという特徴があるため，打楽器音と同様にコード認識に悪影響を与える成分である．本研究では，その悪影響を取り除くために，ロバスト主成分分析に基づく歌声・伴奏音分離とメディアンフィルタに基づく調波音・打楽器音分離を用いて音楽音響信号をあらかじめ歌声・伴奏音・打楽器音に分離する．分離された調波伴奏音から抽出したクロマベクトルを音響特徴量とし，調とコードの遷移を表した隠れマルコフモデルに対してビタビ探索を行うことで最適なコード列を求める．実験では，提案法によりコード認識率が 3.5 ポイント上昇することを確認した．

1. はじめに

音楽音響信号のコード認識は，音楽情報処理の分野における重要な課題の一つである [1], [2]．楽曲中のコードは作曲者認識 [3] やジャンル分類 [4] の手がかりとなる上に，楽器を演奏するための楽譜としても利用できる [5] ことがその理由である．コード認識は一般的に，音響特徴量の抽出とその分類で構成される．コード認識によく用いられる音響特徴量として，12 次元クロマベクトル [6] がある．これは，12 のピッチクラス（C，C#，D，…，B）のエネルギー分布を表したものである．特徴量抽出では，このクロマベクトルをフレーム単位やビート単位で計算することが一般的である．特徴量分類には，各コード間の遷移確率と各コードに対する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 [7], [8], [9]．

多くのコード認識の従来法では，混合音から直接クロマベクトルを抽出している．ポピュラー音楽はたいていの場合，メロディー，伴奏，打楽器の 3 つの異なる役割を持ったパートから構成されている．メロディー（歌声やギターソロなど）は，楽曲のメインのパートであり, 一般的に最も音量が大きいがコードの構成音ではない音を含むことも多い．その一方で，伴奏（リズムギターやピアノなど）はメロディより音量は小さいが，和音を構成するパートであり，コードの決定に大きく関与している．打楽器（ドラムなど）は特定の音高を持たないため，コードとの関係性はない．このような特徴から，クロマベクトルは混合音から直接抽出するのではなく，伴奏音音のようなコードに関係するパートのみを用いて抽出するべきである．いくつかの先行研究では，そのような観点からコード認識率を改善させている．Ueda ら [10] は，打楽器音を抑制した音楽音響信号からクロマベクトルを抽出し，コード認識を行っている．Sumi ら [11] は，ベース音高とコードの依存性に着目し，ベース音高軌跡を用いたコード認識手法を提案している．また，音域ごとの重要性の違いに着目したコード認識手法も提案されている．Mauch ら [12] は一般的なクロマベクトルに加えて低音域のみのエネルギーの分布を表したベースクロマを用いてコード認識を行っている．Cho ら [13] は，音楽音響信号をいくつかの周波数帯域に分け，それぞれの帯域でクロマベクトルを計算している．

本稿では，歌声と伴奏音のコードとの関係性の違いに着目したコード認識手法を提案する．具体的には，まずロバスト主成分分析（RPCA）[14] とメディアンフィルタ [15] を用いて，音楽音響信号を歌声，伴奏音，打楽器音の 3 つに分離する．分離された伴奏音から，音響特徴量であるクロマベクトルを抽出する．その際，倍音の影響を取り除くために，ベイジアン非負値行列因子分解（NMF）による各

[1] 京都大学
[a)] smaruo@kuis.kyoto-u.ac.jp
[b)] ikemiya@kuis.kyoto-u.ac.jp
[c)] itoyama@kuis.kyoto-u.ac.jp
[d)] yoshii@kuis.kyoto-u.ac.jp

音高の音量推定 [16] を行う．コード列はコードと調の遷移を表したベイジアン HMM に対して，ビタビ探索を行うことで推定する．

2. 提案手法

本章では，提案手法である歌声・伴奏音・打楽器音分離（VHPSS）を用いたコード認識手法（**図 1**）について述べる．提案手法は，特徴量抽出と特徴量分類の 2 つの部分に大きく分けられる．特徴量抽出部分では，RPCA およびメディアンフィルタに基づく VHPSS [14], [15] を行い，分離された調波伴奏音からクロマベクトルを抽出する．クロマベクトルを抽出する際，倍音による悪影響を取り除くために，ベイジアン NMF を用いた各音高の音量推定を行う [16]．特徴量分類部分では，抽出した音響特徴量を観測とし，調とコードの対を隠れ状態としたベイジアン転調 HMM に対してビタビ探索を行う．

2.1 問題設定

コード認識の目標は入力音楽音響信号からコードラベルの列を得ることである．コード認識は調の推定とともに行われることが多いため，本稿でもコードと同時に調を推定する．本研究では，入力音楽音響信号から抽出されたクロマベクトルの列 $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_T\}$ を対応するコードラベルと調の列 $\hat{\boldsymbol{Z}} = \{\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_T\}$ に変換することを目指す．この変換は，コードアノテーションが付与された音楽音響信号を用いてあらかじめ学習した統計的分類器を用いて行う．音楽音響信号から抽出したクロマベクトルの列を $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$，それに対応するコードラベルと調の列を $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_T\}$ とする．ここで，$\boldsymbol{z}_t = \{z_t^c, z_t^k\}$ である．$\boldsymbol{X}$ と $\boldsymbol{Z}$ はともに学習データとして与えられる．

コードラベルは，ルート音（C, C#, D, …, B）とタイプ（“maj”, “min”）の組み合わせで表現する．また，“no chord” は記号 N を用いて表現する．調も同様に基音（C, C#, D, …, B）とタイプ（“maj”, “min”）の組み合わせで表現する．したがって，コード語彙数 C は 25，調語彙数 K は 24 となる．本稿の目的はコード認識に対する VHPSS の有効性を示すことなので，以下の仮定の下でコード認識の問題に取り組む．

- ビートは [17] を用いてあらかじめ推定する．
- コードの境界は拍（四分音符）または裏拍（八分音符）の位置に現れるものとする．
- “maj” と “min” の 2 つのコードタイプのみ取り扱う．その他のコードタイプ（“maj7” や”dim” など）は [1] を参考に 2 つのコードタイプのどちらかに分類する．

2.2 歌声・伴奏音・打楽器音分離

コード認識の音響特徴量である伴奏音クロマベクトルは，VHPSS を用いることで入力の音楽音響信号から抽出する．**図 2** に示すように，VHPSS は RPCA に基づく歌声分離 [14] とメディアンフィルタに基づく調波音・打楽器音分離（HPSS） [15] からなる．

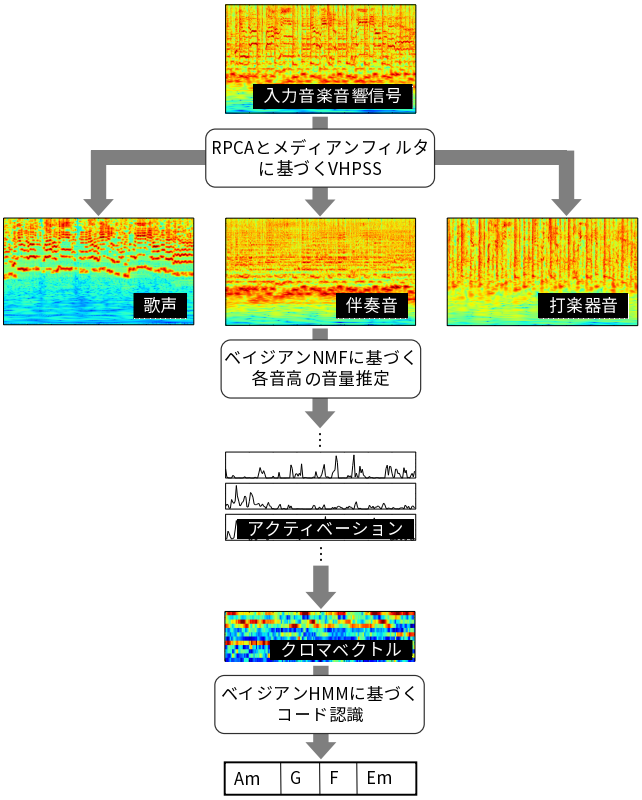


図 1: 提案法の全体像.

2.2.1 歌声分離

入力の音楽音響信号を歌声と伴奏音に分離するために，RPCA に基づく歌声分離 [14] を行う．この手法は歌声分離と歌声 F0 推定の相互依存性を利用している．

まず，RPCA を用いて歌声分離を行う．RPCA は以下のように行列 $\boldsymbol{U}$ を低ランクの行列 $\boldsymbol{L}$ とスパースな行列 $\boldsymbol{S}$ に分解する．

$$\text{minimize}(\|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1), \text{ ただし } \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{U} \quad (1)$$

ここで，$\|\cdot\|_*$ および $\|\cdot\|_1$ はそれぞれ核ノルムと L1 ノルムを表し，λ は $\boldsymbol{L}$ と $\boldsymbol{S}$ のバランスを調節するパラメータである．伴奏音（ギターやピアノなど）は音色や音高が固定されており，基本的に同じ演奏を繰り返す性質があるため，伴奏音のスペクトログラムは低ランクの行列になる．これに対し，歌声は音色や音高が時間によって連続的に大きく変化するため，歌声のスペクトログラムはスパースな行列になる．したがって，RPCA を行うことで，混合音のスペクトログラム $\boldsymbol{U}$ は伴奏音のスペクトログラム $\boldsymbol{L}$ と歌声のスペクトログラム $\boldsymbol{S}$ に分離される．さらに，$\boldsymbol{L}$ の各要素を $\boldsymbol{S}$ の対応する要素と比較することで，時間－周波数バイナリマクス $\boldsymbol{M}_r$ を生成する．このバイナリマスクを混合音のスペクトログラムに適用することで，歌声を抽出する．

続いて，Subharmonic Summation（SHS）[18] およびビタビ探索 [19] を用いて分離された歌声から歌声 F0 軌跡を推定する．歌声 F0 軌跡 $\hat{S}$ は以下の最適経路探索問題を解くことで得られる．

$$\hat{S} = \arg\max_{s_1,\cdots,s_T} \sum_{t=1}^{T-1} \{\log a_t H(t, s_t) + \log T(s_t, s_{t+1})\} \quad (2)$$

ここで t は時間フレームのインデックス，s_t はフレーム t における対数周波数である．$H(t, s_t)$ は SHS によって得られる関数であり，各時間 - 周波数ビンの歌声 F0 らしさを表している．また，$T(s_t, s_{t+1})$ は周波数 s_t から s_{t-1} への歌声 F0 の遷移確率であり，a_t は正規化定数である．ビタビ探索を用いることで，最適な歌声軌跡 $\hat{S}$ を見つけることができる．

最後に，RPCA マスク $\boldsymbol{M}_r$ および調波マスク $\boldsymbol{M}_h$ の 2 つの時間 - 周波数バイナリマスクを用いて，もう一度歌声分離を行う．RPCA マスク $\boldsymbol{M}_r$ は，各時間 - 周波数ビンを歌声か伴奏音のどちらかに分類するマスクであり，RPCA に基づく分離によって既に得られている．調波マスク $\boldsymbol{M}_h$ は，歌声の倍音成分のみを通過させるマスクであり，推定された歌声 F0 軌跡を用いて以下の式で得られる．

$$M_h(t, f) = \begin{cases} 1 & \text{if } \ nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

ここで，F_t はフレーム t における推定された歌声 F0，n は倍音成分のインデックス，w は各倍音成分の周辺でエネルギーを抽出する周波数の幅である．2 つのマスク $\boldsymbol{M}_r$ および $\boldsymbol{M}_h$ を用いることで，以下のように歌声のスペクトログラム $\boldsymbol{V}$ および伴奏のスペクトログラム $\boldsymbol{A}$ が得られる．

$$\boldsymbol{V} = \boldsymbol{M}_r \otimes \boldsymbol{M}_h \otimes \boldsymbol{U} \quad (4)$$

$$\boldsymbol{A} = \boldsymbol{U} - \boldsymbol{V} \quad (5)$$

ここで，$\boldsymbol{U}$ は入力の音楽音響信号のスペクトログラムであり，$\otimes$ は要素積を表す．

2.2.2 調波音・打楽器音分離

メディアンフィルタに基づく HPSS [15] を用いて，歌声分離で得られた伴奏音を調波伴奏音と打楽器音に分離する．この分離手法は，スペクトログラム上で調波音は時間軸方向に滑らかであり，打楽器音は周波数軸方向に滑らかである特徴を利用している．

まず，以下のようにメディアンフィルタを適用することで，伴奏音のスペクトログラム $\boldsymbol{A}$ から調波音を強調したスペクトログラム $\boldsymbol{H}$ と打楽器音を強調したスペクトログラム $\boldsymbol{P}$ を得る．

$$H_f = \mathcal{M}\{A_f, l_{harm}\}, \quad P_t = \mathcal{M}\{A_t, l_{perc}\} \quad (6)$$

ここで，$\mathcal{M}$ はメディアンフィルタ，l_{harm} と l_{perc} はフィルタの長さ，t と f はそれぞれ時間軸上と周波数軸上のインデックスである．次に，得られた 2 つのスペクトログラムを用いて，以下のようにウィーナーフィルタに基づくソフトマスクを生成する．

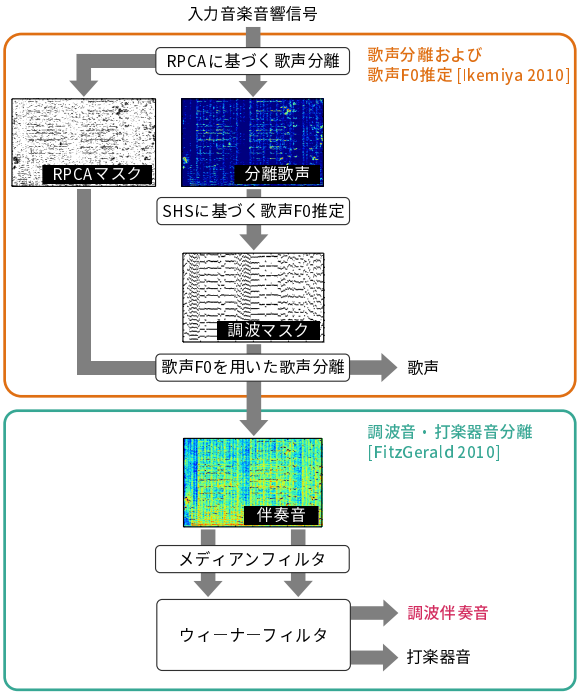


図 2: VHPSS の全体像.

$$M_{Ht,f} = \frac{H_{t,f}^p}{H_{t,f}^p + P_{t,f}^p}, \quad M_{Pt,f} = \frac{P_{t,f}^p}{H_{t,f}^p + P_{t,f}^p} \quad (7)$$

ここで，p は調節可能なパラメータであり，一般的に $p = 2$ のとき良い分離結果が得られる．

最後に，生成したマスクをを用いて，以下のように調波伴奏音と打楽器音の複素スペクトログラムを得る．

$$\hat{\boldsymbol{H}} = \hat{\boldsymbol{S}} \otimes \boldsymbol{M}_H, \quad \hat{\boldsymbol{P}} = \hat{\boldsymbol{S}} \otimes \boldsymbol{M}_P \quad (8)$$

ここで，$\otimes$ は要素積，$\hat{*}$ は複素スペクトログラムを表す．2 つのスペクトログラム $\hat{\boldsymbol{H}}$ と $\hat{\boldsymbol{P}}$ それぞれを逆短時間フーリエ変換することで，分離された調波伴奏音と打楽器音が得られる．得られた調波伴奏音から抽出したクロマベクトルをコード認識の際に音響特徴量として使用する．

2.3 クロマベクトルの抽出

コード認識のための音響特徴量には，伴奏音のクロマベクトルを使用する．クロマベクトルは，倍音を除去するためにベイジアン NMF による各音高の音量推定 [16] を行ってから抽出する．具体的には，ベイジアン NMF によって得られる各音高のアクティベーションを足し合わせることでクロマベクトルを計算する．

NMF は音楽音響信号の多重音解析や音源分離によく用いられる [20], [21]．NMF によって，非負値の行列 $\boldsymbol{Y} \in \mathbb{R}^{F \times T}$ は $\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{H}$ となるように 2 つの非負値の行列 $\boldsymbol{W} \in \mathbb{R}^{F \times R}$

と $\boldsymbol{H} \in \mathbb{R}^{R \times T}$ に分解される．$\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{H}$ ここで，$\boldsymbol{Y}$ を音楽音響信号のスペクトログラムとすると，$\boldsymbol{W}$ は F 個の周波数ビンを持つ R 個の基底スペクトル，$\boldsymbol{H}$ は T 個の時間フレームを持つ R 個のアクティベーションとなる．

ベイジアン NMF では，観測スペクトログラム $\boldsymbol{Y}$ について以下の生成過程を仮定する．

$$p(Y_{mn}) = \mathcal{P}\left(Y_{tf} \middle| \sum_r w_{rf} h_{rt}\right) \tag{9}$$

$$p(w_{rm}) = \mathcal{G}(w_{rm}|a^w_{rm}, b^w_{rm}) \tag{10}$$

$$p(h_{rn}) = \mathcal{G}(h_{rn}|a^h_{rn}, b^h_{rn}) \tag{11}$$

ここで，$\mathcal{P}$ と $\mathcal{G}$ はそれぞれポアソン分布とガンマ分布を表す．変分ベイズ（VB) 法を用いることで，$\boldsymbol{H}$ と $\boldsymbol{W}$ の事後分布は以下のように事前分布と同じくガンマ分布となる．

$$q(w_{rf}) = \mathcal{G}(w_{rf}|a^w_{rf} + \textstyle\sum_t \lambda_{rtf} Y_{tf}, b^w_{rf} + \sum_t \mathbb{E}[h_{rt}]) \tag{12}$$

$$q(h_{rt}) = \mathcal{G}(h_{rt}|a^h_{rt} + \textstyle\sum_f \lambda_{rtf} Y_{tf}, b^h_{rt} + \sum_f \mathbb{E}[w_{rf}]) \tag{13}$$

ここで，$\lambda_{rtf} \propto \exp(\mathbb{E}[\log w_{rf}] + \mathbb{E}[\log h_{rt}])$ は $\sum_r \lambda_{rtf} = 1$ となる補助変数である．

音高ごとのアクティベーションを得るために，各基底スペクトルが半音階の調波構造を持つように設定する．本稿では，A0 から C8（MIDI ノートナンバーの 21 から 108）の 88 個の音高を基底スペクトルに割り当てる．調波構造は混合ガウスモデル（GMM）を用いて表し，倍音成分の大きさは 0.6 倍ずつ指数的に減少していくものとする．具体的には，ハイパーパラメータ $\boldsymbol{a}^w$ を調波構造に比例するように設定し，$\boldsymbol{b}^w = \boldsymbol{1}$ とすることで，基底スペクトルが調波構造となるようにする．ハイパーパラメータ $\boldsymbol{a}^h$ と $\boldsymbol{b}^h$ によりアクティベーションに事前知識を導入することもできるが，本研究では使用しない．すなわち，$\boldsymbol{a}^h$ と $\boldsymbol{b}^h$ はどちらも $\boldsymbol{1}$ とする．

ベイジアン NMF によって得られた各音高のアクティベーションから，各フレーム t での 12 次元クロマベクトル $\boldsymbol{x}_t$ を計算する．具体的には，$\boldsymbol{x}_t$ はアクティベーションの値をオクターブごとに足し合わせることで計算される．ただし，本稿では以下の式のように足し合わせる際に音高ごとに異なる重みを与える．

$$x_n t = \sum_{p:p \equiv n (\mathrm{mod}\ 12)} w_p \cdot h_{pt} \tag{14}$$

ここで，$n \in [0, 11]$ であり，h_{pt} は音高 p，フレーム t のアクティベーションである．w_p は音高ごとの重みであり，幅 88 のハニング窓を使用する．

フレーム単位で得られたクロマベクトルは，半拍ごとに平均を取ることで半拍単位のクロマベクトルに変換する．最後に，クロマベクトルの各次元の値を，平均が 0，分散が 1 となるように正規化する．

2.4 ベイジアン HMM に基づくコード認識

ここでは，ベイジアン HMM を用いてコードと調の列を推定する方法について述べる．コード認識で用いる HMM は，12 次元の GMM を用いて音響特徴量の生成過程を表現している．この HMM に対してビタビ探索を行うことで，最適なコード列を得ることができる．

本稿では，[16] で提案されたベイジアン調依存 HMM と同様の方法でクロマベクトルを分類できるベイジアン転調 HMM を使用する．ベイジアン転調 HMM がベイジアン調依存 HMM と異なるのは，潜在変数 $\boldsymbol{Z}$ がコードと調の対になっている点である．音響特徴量を分類するために，学習データ $\boldsymbol{X}$（特徴量）と $\boldsymbol{Z}$（コードと調のアノテーション）を用いて，ベイジアン転調 HMM の学習を行う．ガウス分布の混合比，平均，精度をそれぞれ $\boldsymbol{\pi}$，$\boldsymbol{\mu}$，$\boldsymbol{\Lambda}$，遷移確率を $\boldsymbol{\phi}$ とする．ベイズ則に従うと，モデルパラメータの事後分布 $p(\boldsymbol{\phi}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z})$ は $p(\boldsymbol{\phi}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z}) p(\boldsymbol{\phi}|\boldsymbol{Z})$ と分解できるため．2 つの事後分布 $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z})$ および $p(\boldsymbol{\phi}|\boldsymbol{Z})$ を学習する．

学習データ $\boldsymbol{X}$ および $\boldsymbol{Z}$ を最大限に使用するために，ピッチクラスの循環性を利用してモデルの学習を行う [10].

- $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}, \boldsymbol{Z})$ **の学習：**クロマベクトル $\boldsymbol{x}_t$ を巡回シフトすることで各コード z^c_t のルート音を C にする．これにより，2 つのコードタイプ（C major と C minor）についてのみ GMM を学習すれば良い．他の 22 個の GMM はガウシアンのパラメータをシフトすることで得られる．

- $p(\boldsymbol{\phi}|\boldsymbol{Z})$ **の学習：**遷移前の調 z^k_{t-1} の基音を C にシフトし遷移後の調 z^k_t の基音も同じ分だけシフトする．それに伴い．遷移前のコード z^c_{t-1} のルート音，遷移後のコード z^c_t のルート音も同じ分だけシフトする．（例えば，調が E♭ major，コードが C minor の状態から，調が C minor，コードが C minor の状態に遷移する場合は，調が C major，コードが A minor の状態から，調が A minor，コードが A minor の状態に遷移したものとして取り扱う．）これにより，2 つの調（C major と C minor）からの遷移についてのみ学習すれば良い．他の 22 の調からの遷移確率は $\boldsymbol{\phi}$ の要素を並べ替えることで得られる．

学習した HMM に対してビタビアルゴリズムを用いることにより，最適な $\hat{\boldsymbol{Z}}$ を探索して入力音楽音響信号のコード列を推定する．本稿では，外れ値に対して頑健な *predictive* HMM を用いる．これは，以下のような遷移確率と出力分布で構成される．

$$p(\hat{\boldsymbol{z}}_t = \{c', k'\} | \hat{\boldsymbol{z}}_{t-1} = \{c, k\}, \boldsymbol{Z}) = \mathbb{E}[\phi_{ckc'k'}] \tag{15}$$

$$p(\hat{\boldsymbol{x}}_t | \hat{\boldsymbol{z}}_t = \{c, k\}, \boldsymbol{X}, \boldsymbol{Z}) = \sum_{l=1}^{L} \mathbb{E}[\pi_{cl}] \mathrm{St}(\hat{\boldsymbol{x}}_t | \boldsymbol{u}_{cl}, \boldsymbol{V}_{cl}, \nu_{cl} - D) \tag{16}$$

ここで，St はスチューデントの t 分布，D は $\hat{\boldsymbol{x}}_t$ の次元，$\boldsymbol{u}$

と ν はハイパーパラメータである．$\boldsymbol{V}_{kl}$ は以下の式で与えられる．

$$\boldsymbol{V}_{cl} = \frac{(\nu_{cl} - D)\beta_{cl}}{1 + \beta_{cl}} \boldsymbol{W}_{cl}, \tag{17}$$

ここで，β と $\boldsymbol{W}$ はハイパーパラメータである．

3. 評価実験

提案手法の有効性を示すために，実際の音楽音響信号に対してコード認識を行った．

3.1 実験条件

コード認識の評価には The Beatles のデータセット [22] の 179 曲を使用し，全楽曲を無作為に 10 個のグループに分けて 10-fold Cross Validation を行った．すべての楽曲のサンプリングレートは 16 kHz であり，VHPSS を行う際の STFT は窓幅 128 ms，ステップ幅 10 ms とした．SHS の倍音の数は 10，RPCA のパラメータ k（[23] に記載）は 1.0，式 (3) におけるマスクの幅 w は 30 Hz とした．歌声 F0 の探索範囲は 120 Hz から 720 Hz とした．各音高の音量推定を行う際のスペクトログラムは定 Q 変換により計算し，周波数の間隔は 20 cent，ステップ幅は 50 ms とした．ベイジアン HMM のハイパーパラメータ（[16] に記載）は $L = 32$, $\boldsymbol{\alpha}_0 = \mathbf{1}$, $\boldsymbol{\gamma}_0 = \mathbf{1}$, $\boldsymbol{u}_0 = \mathbf{0}$, $\boldsymbol{W}_0 = \boldsymbol{I}$, $\beta_0 = \nu_0 = 12$ とした．コード認識率は，評価データ全体に対して認識結果が正解であった区間の割合として計算した．

コード認識の評価には以下の 4 種類の音響特徴量を使用した．

Original 元の音楽音響信号から抽出したクロマベクトル

HPSS HPSS により打楽器音を取り除いた音楽音響信号から抽出したクロマベクトル

VHPSS VHPSS により歌声と打楽器音を取り除いた音楽音響信号から抽出したクロマベクトル

また，クロマベクトルは以下の 2 種類の方法で抽出した．

倍音除去なし ベイジアン NMF の基底スペクトルは各音高の F0 とする．すなわち，倍音は除去されない．

倍音除去あり ベイジアン NMF のの基底スペクトルは調波構造とする．すなわち，倍音が除去される．

3.2 実験結果

表 1 に実験結果を示す．VHPSS を用いることによりコード認識率が改善されたことから，提案法が有効であることが示された．

図 3a は HPSS または VHPSS を行った場合のクロマベクトルと認識されたコードラベルの一例である．HPSS を行った場合のクロマベクトルは，コードの特徴があまり明確に現れていない．これは，音量の大きい歌声によって伴奏音が埋もれてしまっているためである．そのため，枠で囲んだ部分では認識されたコードラベルが間違いとなっている．一方で，VHPSS を行った場合のクロマベクトルは，歌声が除去されたことにより伴奏音の音高分布がより鮮明となり，コードの特徴が明確に現れている．これにより，枠で囲んだ部分の認識されたコードラベルが正解となっている．

表 1: コード認識の実験結果.

	Original	HPSS	VHPSS
倍音除去なし	72.4	73.7	75.6
倍音除去あり	73.8	74.7	77.3

また，NMF に基づく各音高の音量推定によって倍音成分を除去することで，コード認識率は更に改善した．**図 3b** は NMF の基底スペクトルを F0 のみとした場合，調波構造とした場合それぞれのクロマベクトルと認識されたコードラベルである．これらのクロマベクトルは VHPSS を行って得られた調波伴奏音から抽出されたものである．基底スペクトルが F0 のみの場合，クロマベクトルに倍音成分が現れ，コードの特徴が不鮮明である．そのため，枠で囲んだ部分では認識されたコードラベルが間違いとなっている．一方で，基底スペクトルが調波構造の場合は，倍音成分が除去されることにより，クロマベクトルにコードの特徴が鮮明に現れている．そのため，枠で囲んだ部分の認識されたコードラベルが正解となっている．

提案手法によって得られたコード認識率のうち最大であったのは 77.3 %で，基底スペクトルを調波構造とし，音響特徴量として調波伴奏音のクロマベクトルを用いた場合であった．

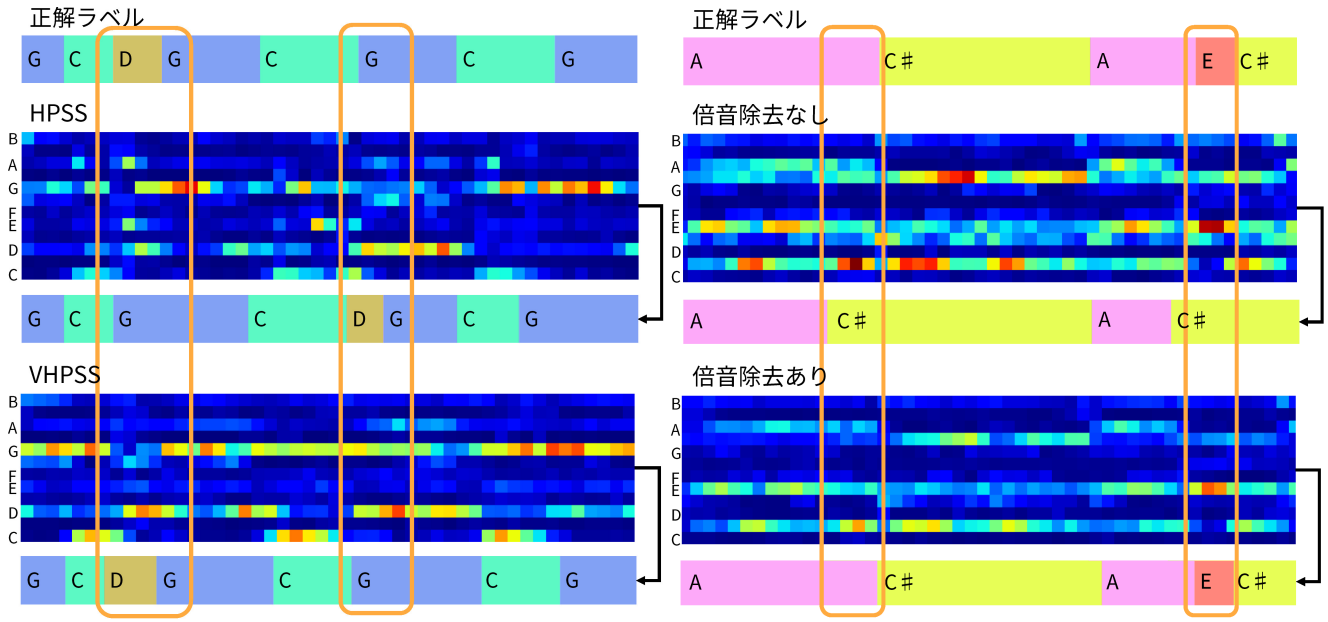
4. おわりに

本稿では，歌声・伴奏音・打楽器音分離（VHPSS）に基づくコード認識手法を提案した．実験結果により，提案手法の有効性が示された．VHPSS により分離された調波伴奏音のみを用いてコード認識を行うことで認識率が改善された．更にコード認識率を改善するために，今後はコードに大きく依存しているベース音も用いてコード認識を行う予定である．提案手法では，既存のビートトラッキング手法を用いてビートをあらかじめ推測し，コードの境界は表拍または裏拍に現れると仮定した．この制約を取り払うために，コードラベルと同時にビートの位置とコードの境界を推定するようなモデルを提案することで，更にコード認識率を改善できると考えられる．

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] Mauch, M.: Automatic chord transcription from audio using computational models of musical context, PhD



(a) HPSS と VHPSS の比較　　(b) 倍音除去なしと除去ありの比較

図 3: クロマベクトルと認識結果の例.

Thesis, School of Electronic Engineering and Computer Science Queen Mary, University of London (2010).

[2] Harte, C.: Towards automatic extraction of harmony information from music signals, PhD Thesis, Department of Electronic Engineering, Queen Mary, University of London (2010).

[3] Ogihara, M. and Li, T.: N-Gram Chord Profiles for Composer Style Representation, *ISMIR 2008*, pp. 671–676 (2008).

[4] Pérez-Sancho, C., Rizo, D. and Iñesta, J. M.: Genre classification using chords and stochastic language models, *Connection science*, Vol. 21, No. 2-3, pp. 145–159 (2009).

[5] Weil, J., Sikora, T., Durrieu, J.-L. and Richard, G.: Automatic generation of lead sheets from polyphonic music signals, *ISMIR 2009*, pp. 603–608 (2009).

[6] Fujishima, T.: Realtime chord recognition of musical sound: A system using common lisp music, *ICMC 1999*, pp. 464–467 (1999).

[7] Sheh, A. and Ellis, D. P. W.: Chord segmentation and recognition using EM-trained hidden Markov models, *ISMIR 2003*, pp. 185–191 (2003).

[8] Lee, K. and Slaney, M.: A Unified System for Chord Transcription and Key Extraction Using Hidden Markov Models, *ISMIR 2007*, Citeseer, pp. 245–250 (2007).

[9] Chen, R., Shen, W., Srinivasamurthy, A. and Chordia, P.: Chord Recognition Using Duration-explicit Hidden Markov Models, *ISMIR 2012*, pp. 445–450 (2012).

[10] Ueda, Y., Uchiyama, Y., Nishimoto, T., Ono, N. and Sagayama, S.: HMM-based approach for automatic chord detection using refined acoustic features, *ICASSP 2010*, IEEE, pp. 5518–5521 (2010).

[11] Sumi, K., Itoyama, K., Yoshii, K., Komatani, K., Ogata, T. and Okuno, H. G.: Automatic Chord Recognition Based on Probabilistic Integration of Chord Transition and Bass Pitch Estimation., *ISMIR 2008*, pp. 39–44 (2008).

[12] Mauch, M. and Dixon, S.: Approximate Note Transcription for the Improved Identification of Difficult Chords., *ISMIR 2010*, pp. 135–140 (2010).

[13] Cho, T. and Bello, J. P.: MIREX 2013: Large vocabulary chord recognition system using multi-band feature and a multi-stream HMM, *MIREX 2013* (2013).

[14] Ikemiya, Y., Yoshii, K. and Itoyama, K.: Singing voice analysis and editing based on mutually dependent F0 estimation and source separation, *ICASSP 2015*, pp. 574–578 (2015).

[15] FitzGerald, D.: Harmonic/percussive separation using median filtering, *DAFx-10* (2010).

[16] Maruo, S., Yoshii, K., Itoyama, K., Mauch, M. and Goto, M.: A feedback framework for improved chord recognition base on NMF-based approximate note transcription, *ICASSP 2015* (2015).

[17] Dixon, S.: Evaluation of the audio beat tracking system beatroot, *Journal of New Music Research*, Vol. 36, No. 1, pp. 39–50 (2007).

[18] Hermes, D. J.: Measurement of pitch by subharmonic summation, *The journal of the acoustical society of America*, Vol. 83, No. 1, pp. 257–264 (1988).

[19] Viterbi, A. J.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *Information Theory, IEEE Trans. on*, Vol. 13, No. 2, pp. 260–269 (1967).

[20] Raczyński, S. A., Ono, N. and Sagayama, S.: Multipitch analysis with harmonic nonnegative matrix approximation, *ISMIR 2007*, pp. 381–386 (2007).

[21] Vincent, E., Bertin, N. and Badeau, R.: Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription, *ICASSP 2008*, pp. 109–112 (2008).

[22] Mauch, M., Cannam, C., Davies, M., Dixon, S., Harte, C., Kolozali, S., Tidhar, D. and Sandler, M.: OMRAS2 metadata project 2009, *ISMIR 2009* (2009).

[23] Huang, P. S., Chen, S. D., Smaragdis, P. and Johnson, M. H.: Singing-voice separation from monaural recordings using robust principal component analysis, *ICASSP 2012*, pp. 57–60 (2012).