Vol.2015-AVM-91 No.28
2015/12/4

Harmonic/Percussive Sound Separation を前処理とした和音認識の性能調査

安田龍一郎† 大石皓太郎† 植村あい子† 甲藤二郎†

概要：本稿では，Harmonic/Percussive Sound Separation (HPSS)を前処理として和音認識を行う際の，認識結果への影響を調査する．HPSS を前処理として和音認識を行うと認識精度が向上することが従来研究により明らかになっている．HPSS を行う際の STFT のフレーム長，シフト長を変化させると，分離結果に違いが見られるため，本稿では，HPSS のフレーム長，シフト長に関し，複数条件のもと分離結果の違いが認識結果に与える影響を調査する．和音認識は従来手法であるクロマベクトルと HMM を用い，楽曲データセット 307 曲を用いて評価を行った．

キーワード：和音認識，Harmonic/Percussive Sound Separation，クロマベクトル，HMM

Performance survey of Chord recognition using Harmonic/Percussive Sound Separation as Preprocessor

RYUICHIRO YASUDA† KOTARO OISHI†
AIKO UEMURA† JIRO KATTO†

Abstract: In this paper, we examine the influence on chord recognition results using Harmonic/Percussive Sound Separation (HPSS) as a preprocessor. It has been known that the recognition accuracy is improved when HPSS is applied to chord recognition tasks. However, when changing frame length and shift length of STFT to calculate HPSS, different results are observed in time-frequency separation. Therefore, in this paper, we try various conditions about frame length and shift length, and examine their influences on chord recognition results. We use a chroma vector and HMM which have been commonly used in chord recognition tasks, and carry out experiments by using 307 songs as music dataset.

Keywords: Chord Recognition, Harmonic/Percussive Sound Separation, Chroma Vector, HMM

1. はじめに

近年のコンピュータの性能の向上により，音楽音響信号を容易に扱うことができるようになり，また同時にこれらの解析はますます重要になってきている．大容量オーディオプレイヤ，更にはスマートフォンの普及や，ネットワークを利用した楽曲配信サービスの発展により，ユーザが音楽に触れる手段は増加し，多種多様なニーズに合わせた楽曲推薦，楽曲検索が求められている．

和音，つまりハーモニーは，メロディ，リズムと並び，音楽の三大要素とされ，音楽を構成する重要な要素である．和音進行から調性や楽曲構造を判断することができることから，和音は他の音楽要素の解析の手掛かりとしても注目され，楽曲推薦，楽曲検索のためにも利用できる．また，本分野においての目標の 1 つである実音響信号から楽譜を起こす作業，通称「耳コピ」の自動化実現のためには，和音認識，音源同定，音源分離，音高推定，テンポ推定など，多くの技術を必要とする．中でも和音認識は特に重要である．例えば人の手で採譜をするときに和音を手掛かりに 1 つ 1 つ音を求めていくように，自動採譜においても和音認識を前処理として利用することが期待される．

和音認識の従来研究は，時系列パターン問題として扱うことが多く，音響信号を入力し，学習・認識を経て，和音系列を出力させるものが主流である．具体的には，既知の音響信号と和音系列から識別器を学習し，認識時に入力音響信号の和音系列を出力する．和音系列はマルコフ過程であると仮定し，和音特徴量を観測データとして Hidden Markov Model(HMM)を用いて和音系列を求める[1]. 和音特徴量は，クロマベクトル[2]がよく用いられる．クロマベクトルは，数オクターブで演奏されたものでも，構成音が同じであれば同一和音と認識されるという性質に基づいている．

ポピュラー音楽はベースやギターなどの和音を構成する調波音の他に，ドラムセットなどの打楽器の様な和音に関係のない非調波音も含まれている．多くの従来研究では，これらの混合音からクロマベクトルを算出している．しかし和音の認識においては，和音を構成する音のみで，またはそれらを強調して，クロマベクトルを算出することが望ましい．この点に着目して和音認識の性能向上を目指している先行研究はいくつかあり，上田ら[3]は音響信号の調波音を強調し，チューニングをしてからクロマベクトルを算出している．また，須見ら[4]は，ベース音高の和音認識へ

†早稲田大学大学院基幹理工学研究科
Graduate School of Fundamental Science and Engineering, Waseda University.

の影響が大きいことから，ベースラインと和音進行との相関に着目した和音認識を行っている．

本稿では，和音認識の前処理として Harmonic/Percussive Sound Separation (HPSS) [5]を行って調波音と非調波音を分離し，和音構成音の含まれる調波音からクロマベクトルを算出して，多変量単一正規分布を用いた HMM により和音認識を行う．このとき，HPSS を行う際の短時間フーリエ変換(STFT)のフレーム長，シフト長を変化させると，分離結果に違いが見られる．この違いが認識結果に与える影響を調査する．

2. Harmonic/Percussive Sound Separation

HPSS とは，スペクトログラムにおいて，調波音成分は時間軸方向に，非調波音成分は周波数軸方向に連続性が強いということに着目し，それらを分離する手法である．
調波音のスペクトログラムを H，非調波音のスペクトログラムを P，時間インデックスを t，周波数インデックスを k とすると，式(1)(2)(3)で表せる． w_H, w_P は重み係数で，$w_H = w_P = 1.0$ である．

$$|H_{t,k}| = \frac{w_H^2\left(|H_{t+1,k}| + |H_{t-1,k}|\right)^2 |W_{t,k}|}{w_H^2\left(|H_{t+1,k}| + |H_{t-1,k}|\right)^2 + w_P^2\left(|P_{t,k+1}| + |H_{t,k-1}|\right)^2} \quad (1)$$

$$|P_{t,k}| = \frac{w_P^2\left(|P_{t,k+1}| + |P_{t,k-1}|\right)^2 |W_{t,k}|}{w_H^2\left(|H_{t+1,k}| + |H_{t-1,k}|\right)^2 + w_P^2\left(|P_{t,k+1}| + |H_{t,k-1}|\right)^2} \quad (2)$$

ただし，

$$|W_{t,k}| = |H_{t,k}| + |P_{t,k}| \quad (3)$$

分離結果は，HPSS を行う際の STFT のフレーム長，シフト長によって変化する．これは，フーリエ変換の不確定性原理に関係し，例えばフレーム長が長いと周波数分解能が高く時間分解能が低い．周波数分解能が高いと僅かな周波数変化も観測でき周波数軸方向の連続性に敏感になる．しかし時間分解能が低いと，長時間一定の周波数を維持していないと時間軸方向に連続性があると判断できない．つまり，フレーム長が長いほど調波音成分への分離の条件は厳しくなる．この逆も同様であり、このことから目的に合わせたパラメータ設定が重要となる。

3. クロマベクトル

本稿のクロマベクトルは Ellis らの手法[6]に基づき算出する．クロマベクトルは周波数パワースペクトルを特定のピッチクラスに振り分け，半音階 1 オクターブ分に相当する 12 個の中で何の音がどの程度含まれているかを表すものである．

クロマベクトル $v(k,t)$ 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．

$$v(k,t) = \log\left(\sum_{i=0}^{I-1} X(12i+k,t)\right) \qquad k = 0 \cdots 11 \quad (4)$$

ただし，$X(i,t)$ はスペクトルの周波数 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．

本稿では, まず楽曲を 22050Hz にダウンサンプリングし，フレーム長 4410 点（0.2s），シフト長 2205 点（0.1s）で STFT を行い算出する．

4. 和音認識

和音認識をするにあたり，単一フレームの特徴量から和音を推定することは難しい．これは，和声音の省略や，非和声音の混入などがあるためである．また，和音はある程度の時間は遷移せず，和音の遷移には偏りがあるという性質をもつ．そのため各フレームの和音認識は，それらの特徴量だけでなく，その前後のフレームの特徴量も用いることが可能である．現在フレームの和音は，その前フレームの和音の影響を受けると考え，現在フレームの N-1 フレーム前までの和音に依存する N-gram モデルとし，現在フレームの特徴量は隠れ状態の和音から確率的に出力されるとする．本稿では直前フレームのみ影響するとして，2-gram モデルとする．

和音はルート音が異なっていても同種類の和音であれば，それぞれの和声音の間隔は等しく，major はルート音とその長 3 度と完全 5 度で構成され，minor はルート音とその短 3 度と完全 5 度で構成される．この性質から，多変量単一正規分布を求める際には， Cmajor のモデル生成に Cmajor だけでなく C#major から Bmajor までの特徴量も，音をずらして Cmajor とするといったように, ルートをずらしてルートの異なるすべての同種類の和音の特徴量を考慮することで学習データ数を増やしている. Cminor においても同様に求める．そして，Cmajor と Cminor のモデルをそれぞれシフトさせることで全ルートの和音のモデルを求めている．

5. 実験

5.1 実験条件

前処理として HPSS を行い，多変量単一正規分布を用いて出力確率を算出し，HMM により和音認識を行う．処理の流れを図 1 に示す. そして, HPSS を行う際の STFT のフレーム長，シフト長を下記条件のもと変化させ，分離結果の違いが認識結果に与える影響を調査する．

・フレーム長
512, 1024, 2048, 4096, 8192, 16384,32768 点
・シフト長
条件 1：各フレーム長の 1/2
条件 2：全フレーム長の条件においてシフト長 512 点
条件 3：全フレーム長の条件においてシフト長 4096 点

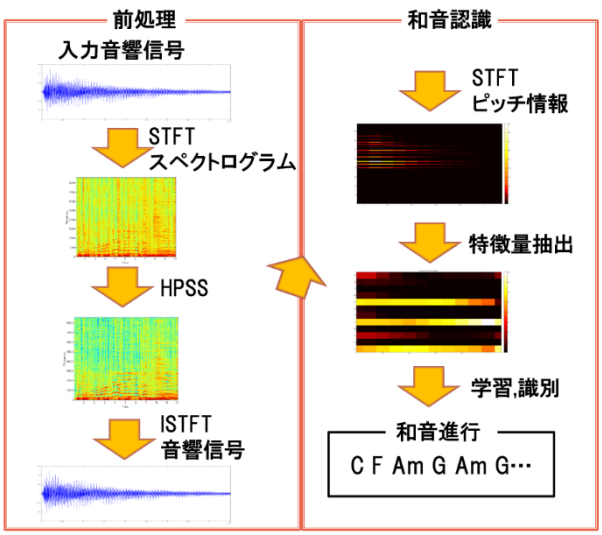

図 1　処理の流れ

ただし，フレーム長がシフト長以下の条件については調査していない．

和音認識に使用する楽曲は The Beatles（180 曲），Queen（20 曲），Carole King（7 曲），RWC（100 曲）計 307 曲の wav 音源 [7][8][9]であり，ビットレート 1411kbps，サンプリング周波数 44100Hz，量子化ビット数 16bit である．扱う和音は major と minor の 24 種類である．実際には楽曲中に major と minor 以外の和音もあるが便宜上それらは major と minor に振り分けている．例えば，sus4 や maj7 は major に，min7 や dim は minor に分類する。全曲の 1/4 を評価データ，残り 3/4 を学習データとして交差確認法により和音認識率を求める．

また，和音認識の正答率は式(5)のように定める．

$$\text{正解和音認識率} = \frac{\text{正しく出力されたフレーム}}{\text{全フレーム数}} \quad (5)$$

5.2 実験結果

HPSS を行う際の STFT のフレーム長, シフト長を各条件にしたときの 307 曲の和音認識率平均の変化を図 2 に示す．フレーム長 512,1024 点のときの全条件とフレーム長 32768 点シフト長 16384 点の条件を除き，HPSS を和音認識の前処理とすることで和音認識率は向上した．中でもフレーム長 16384 点シフト長 4096 点としたときの和音認識率は 63.49%となり，前処理無しの条件と比較して 6.49%向上した．しかし，フレーム長 512 点シフト長 256 点としたときの和音認識率は 50.34%となり，前処理無しの条件と比較して 6.66%低下した．

まず 5.1 の条件①において，シフト長がフレーム長の 1/2 だと，フレーム長が長くなるにつれて時間分解能が著しく低下し，フレーム長 32768 点シフト長 16384 点の条件では認識率が前処理無しの場合よりも低下した．そこで条件②シフト長 512 点に設定変更した．すると，フレーム長 32768

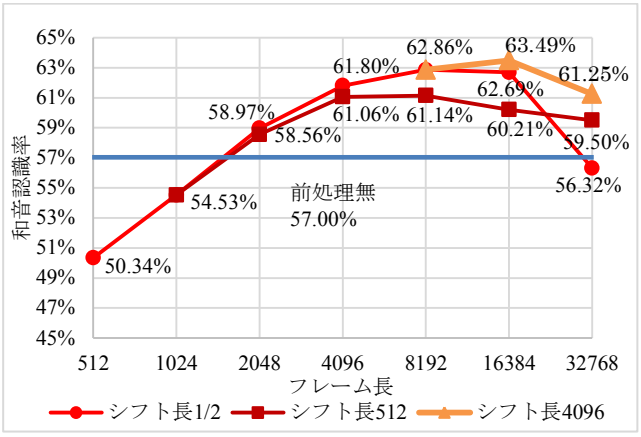

図 2　和音認識率平均の変化

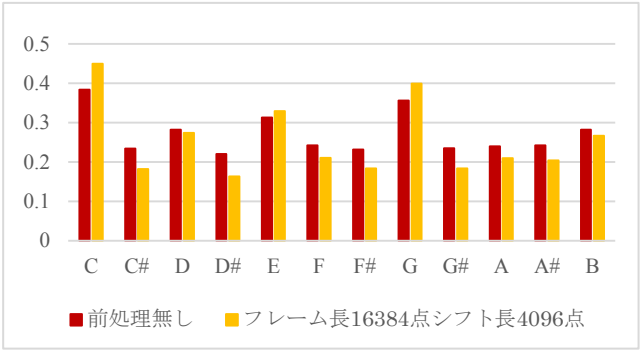

図 3　Cmajor の正規分布の平均の比較

点においては認識率が向上したが，フレーム長が 32768 点より短い条件においては低下した．さらに条件③シフト長 4096 点に変更すると，フレーム長 16384 点 32768 点，両方の認識率が向上した．よってこの結果から，シフト長は長すぎず短すぎないことが好ましいと言える．

6. 考察

6.1 楽曲全体における HPSS の和音認識への影響

HPSS を行う際の STFT のフレーム長 16384 点シフト長 4096 点の条件のときの Cmajor の正規分布の平均を図 3 に示す．Cmajor の和声音 C，E，G が強調され，他の非和声音が抑制されていることが見て取れる．この結果により認識率が向上したと考えられる．

フレーム長，シフト長を短くしたときの和音認識率が低下した原因を考える。まず，表 1 に示す各周波数の sin 波の重ね合わせによってサンプリング周波数 44100Hz の Cmajor の合成音を作成する．その音源にフレーム長 512 点シフト長 256 点の条件と，フレーム長 32678 点シフト長 16384 点の条件のもと HPSS をそれぞれ行う．分離結果として調波音成分を図 4 に示す．(ただし，図 4 に示すスペクトログラムを作成するためのフレーム長シフト長はそれぞれ 16384 点 512 点である．)合成音は完全に調波音のみで構成されているため，HPSS を行ってもほとんど変化がないはずだが，図 4 を見て分かる通り，ノイズが生じてしまっ

表 1　Cmajor の周波数

音名	C1	C3	E4	G4
周波数[Hz]	65.75	261.75	329.75	392.25

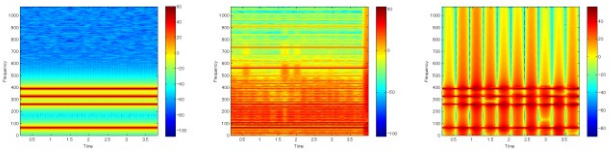

図 4　Cmajor 合成音の HPSS 処理結果　(左:処理前, 中央: フレーム長 512 点シフト長 256 点の条件で HPSS, 右: フレーム長 32678 点シフト長 16384 点の条件で HPSS)

ている.

ここで，スペクトログラムの周波数 bin 幅 Δf [Hz]は，フレーム長 n [点]，サンプリング周波数 Fs [Hz]によって式(6)のように定まる.

$$\Delta f = Fs/n \tag{6}$$

つまりサンプリング周波数 44100Hz，フレーム長 512 点の場合 Δf は約 86Hz となり，C3 と E4 と G4 の周波数差よりも大きく，単一周波数 bin に 2 音含まれてしまう．これによりノイズが発生したとされる．楽曲においても同様の理由のノイズによって和音認識精度が低下したと考えられる.

また，フレーム長 32678 点の場合 Δf は約 1.3Hz となり，非常に幅が狭く，さらに，フレーム長は 0.74s と長いため和音境界があいまいになったことにより精度が落ちたと考えられる.

6.2 各楽曲における HPSS の和音認識への影響

307 曲の和音認識率平均をもとに HPSS の性能を調査したが, 1 曲ごとに認識率の上がり幅, 下がり幅は違うため, どのような曲が HPSS によって和音認識の精度が上がるのか，下がるのかを調査する.

HPSS を行う際の STFT のフレーム長 2048 点シフト長 1024 点の条件では，307 曲中 151 曲の楽曲が前処理によって向上しており, フレーム長 16384 点シフト長 4096 点の条件では，同じく 264 曲の楽曲が前処理によって向上している.

楽曲を聞き比べると、次のような楽曲の和音認識率が向上していることが観察できた。

・フレーム長 2048 点シフト長 1024 点の条件について
(i)ハイハットシンバルを刻んでいる音が大きい楽曲
(ii)ベースラインの動きが安定している楽曲
(iii)ボーカルやギターの音の動きが激しくない楽曲
・フレーム長 16384 点シフト長 4096 点の条件について
(iv)上記の(i)(ii)に加えて、ボーカルの声量が大きい楽曲

HPSS による和音認識精度向上の成功例として、RWC の N096-M07-T06 からの一部抜粋を挙げる。HPSS を行う際の STFT のフレーム長 2048 点シフト長 1024 点の条件(a)と，フレーム長 16384 点シフト長 4096 点の条件(b)におけるスペクトログラムとクロマベクトルを図 5，図 6 に示す．スペクトログラムを見ると，HPSS により，条件(a)では縦筋のパーカッション成分が省かれ，さらに条件(b)では，ボーカルや調波音楽器成分のうねりも省かれていて，まっすぐな横筋のみとなっている．クロマベクトルを見ても細かなノイズの様な物が除去され，和声音であろう音が強調されていることが観察でき,和音認識結果も正しくなっている．この結果, 認識率は表 2 のように 57.80%→63.70%→67.32%へ向上した。

逆に，HPSS の前処理によって和音認識率が低下する楽曲は，上記に当てはまらないものである.
(i)ハイハットシンバルなどのパーカッションの音が小さい
(ii)ベースラインの動きが安定している
(iii)調波音楽器によってハーモニーがわかりやすい
(iv)ボーカルがバラードの様な音に動き小さいメロディ
以上のような楽曲は前処理無しでも，もともと認識率が比較的高く，HPSS の効果が薄い.

認識率が低下する楽曲の例として，The Beatles の Carry That Weight からの一部抜粋を挙げる．条件(a)と，条件(b)におけるスペクトログラムとクロマベクトルを図 7，図 8 に示す.（ただし，図 5 に示すスペクトログラムのフレーム長シフト長は図 4 と同じくそれぞれ 16384 点 512 点である.）また，各条件の和音認識率を表 3 にまとめる．スペクトログラムを見ると，前処理無しでもすでに横筋の調波音成分がくっきりとしている．ここでは主にトランペットがハーモニーを奏でているが，条件(a)ではトランペットの倍音成分が取り除かれ，条件(b)ではアタック部分がぼやけ，2 秒あたりから始まる和音の頭があいまいになっている. また，和声音も薄れてしまっている．図 6 のクロマベクトルも，もともと和声音がくっきりとしていて，HPSS を行ってもあまり変化がなく和音認識結果もほとんど変わらない．よって，もともとクロマベクトルで和声音がはっきりしているような楽曲は HPSS 処理によって和音境界がぼやけて和音認識の精度が落ちることが分かった.

また，認識率が 1%に満たないような楽曲は，HPSS を行っても向上することはほとんどなかった.

表 2　N096-M07-T06 和音認識率

処理無	条件(a)	条件(b)
57.80%	63.70%	67.32%

表 3　Carry That Weight 和音認識率

処理無	条件(a)	条件(b)
91.89%	89.78%	88.41%

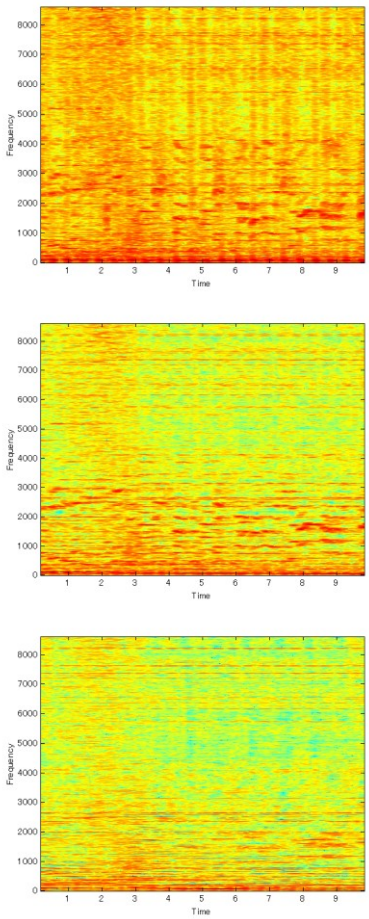

図 5　N096-M07-T06 スペクトログラム(上:前処理無,
中: フレーム長 2048 点シフト長 1024 点の条件(a)で HPSS,
下: フレーム長 16384 点シフト長 4096 点の条件(b)で HPSS)

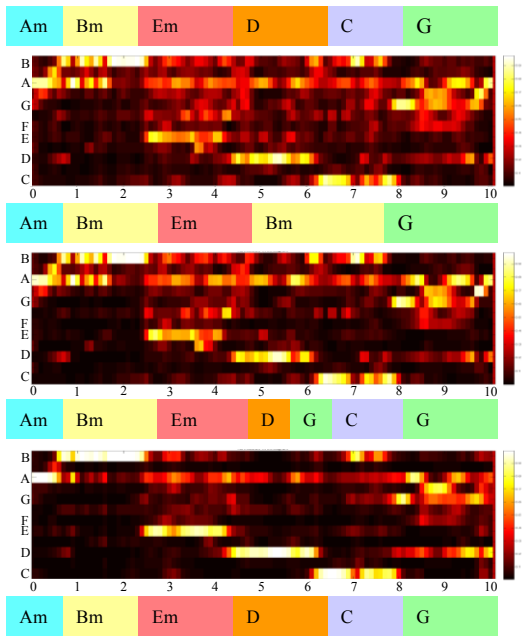

図 6　N096-M07-T06 のクロマベクトル(上:前処理無,
中: フレーム長 2048 点シフト長 1024 点の条件(a)で HPSS,
下: フレーム長 16384 点シフト長 4096 点の条件(b)で HPSS)

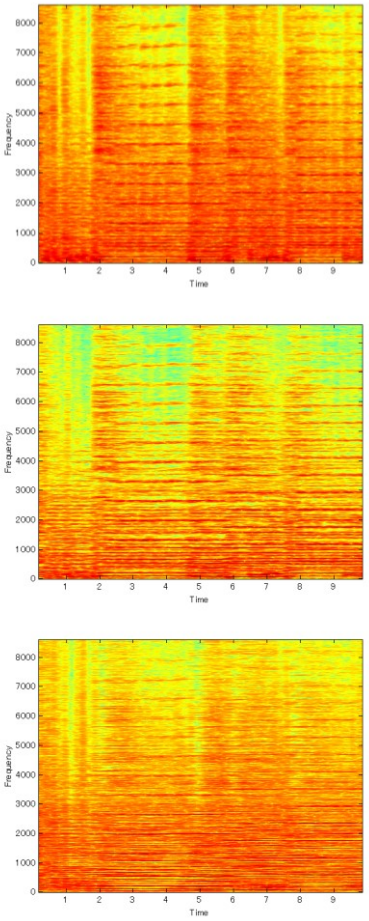

図 7　Carry That Weight のスペクトログラム(上:前処理無,
中: フレーム長 2048 点シフト長 1024 点の条件(a)で HPSS,
下: フレーム長 16384 点シフト長 4096 点の条件(b)で HPSS)

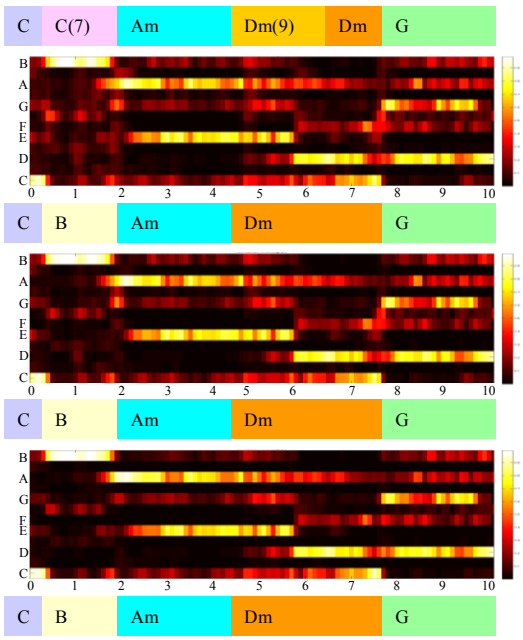

図 8　Carry That Weight のクロマベクトル(上:前処理無,
中: フレーム長 2048 点シフト長 1024 点の条件(a)で HPSS,
下: フレーム長 16384 点シフト長 4096 点の条件(b)で HPSS)

7. おわりに

本稿では，HPSS を前処理として和音認識を行う際の，認識結果への影響を調査した．実験結果より，多くの条件において HPSS を和音認識の前処理とすることで，和音認識率は向上した．特にフレーム長 16384 点シフト長 4096 点としたとき最もその効果が高かった．楽曲によって，HPSS による和音認識率改善の効果は異なり，中には低下するものもあったため，前処理をすべきかすべきでないかを楽曲ごとに自動で判別できるようにしたいと考えている．本稿では単一正規分布を用いたが，今後は混合正規分布を用いた和音認識を行い，HPSS の和音認識への影響を調査する予定である．

参考文献

1) A. Sheh and D. P. Ellis,:Chord segmentation and recognition using EM-trained hidden markov models. Proc. ISMIR, pp.183-189, (2003).
2) T. Fujishima. :Real-time chord recognition of musical sound: A system using common lisp music. Proc. ICMC, pp. 464-467,(1999).
3) 上田 雄, 小野 順貴, 嵯峨山 茂樹: 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).
4) 須見 康平, 糸山 克寿, 吉井 和佳, 駒谷 和範, 尾形 哲也, 奥乃 博: ベース音高を考慮したポピュラー音楽に対する和音進行認識, 情報処理学会第 70 回全国大会(2008).
5) 宮本賢一, 亀岡弘和, 小野順貴, 嵯峨山茂樹: 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).
6) Ellis, Daniel P. W. and Poliner, Graham E.: Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. Proc.ICASPP(2007).
7) Supervised Chord Recognition for Music Audio in Matlab : http://labrosa.ee.columbia.edu/ projects/chords/
8) isophonics : http://isophonics.net/
9) RWC 研究用音楽データベース: http://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-p-j.html