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

上田　雄[†1]　　小野　順貴[†1]　　嵯峨山　茂樹[†1]

本稿では、音楽音響信号から和声進行を認識する問題を扱う。ポピュラー音楽などの音楽音響信号は、一般に打楽器等の非調波音を含み、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・打楽器音分離手法によるクロマベクトルの調波成分の強調と HMM による和音進行のモデリングを用いた和音認識手法を提案してきた。本研究では、さらに楽曲間のチューニングの違いの自動補正により認識性能の向上について検討したので報告する。

Automatic Chord Detection from Musical Acoustic Signals Using Harmonic/Percussive Sound Separation and Tuning Compensation

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

In this paper we discuss a method to automatically detect chord progression from musical acoustic signals. Musical acoustic signals such as popular music contain non-harmonic sounds such as drum sounds, which make it difficult to detect chord progression which depend on only harmonic component. We proposed a method using harmonic-emphasized chroma vector generated by harmonic/percussive sound separation (HPSS) developed in our laboratory. Additionally, we compensate tuning difference among songs automatically. In this paper we evaluate these methods using the Beatles' 180 songs.

†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

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、音楽音響信号からの自動和音認識について扱う。西洋音楽などの調性音楽において、和声進行は音楽の構造における重要な要素の一つであり、実演奏からの自動和音認識は自動採譜、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、和音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。

HMM により和声進行をモデル化した従来研究としては、まず川上らの研究[3] があげられる。川上らの研究では、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されている。Sheh ら[1] は、音楽音響信号の和声進行に対しクロマベクトルと HMM を用いたモデリングを行い、和音の時刻アラインメントを与えずに音響信号と和音進行から出力確率分布と遷移確率分布を、ランダムな初期値から EM アルゴリズムを用いて推定する手法を提案した。Mauch ら[8] はメロディ帯域とバス帯域のクロマベクトルにチューニング手法を適用し、HMM を用いることで 6 つのコードラベルに対し認識を行った。内山らによる手法[6),7)] では本研究室で開発された調波音／打楽器音分離 (Harmonic/Percussive Sound Separation: HPSS)[4] を用いて調波音を強調したクロマベクトルと HMM を用い、MIR の国際コンテストである MIREX 2008[10] の和音認識タスクで第一位を獲得している。

しかしながら、内山らの手法では楽曲間のチューニングは同一であるという仮定の下で認識を行っていた。チューニングとは演奏者が楽器間で音の高さを合わせることであり、共通する音高で合わせさえすればよい。そのため、実際の演奏の録音では、録音された状況や演奏者の好みなどによりチューニングの多少の変動はあり得るため、楽曲ごとのチューニングが一致しているとは限らない。これを無視して、チューニングが一定と仮定してクロマベクトルを求めると、和声の情報がクロマベクトルに明瞭に反映されない可能性がある。こうした問題に対処するため、本研究では内山らの手法に加え、クロマベクトルを求める際にチューニングのずれを自動的に補正する手法を検討し、和音認識性能によりその有効性を評価する。

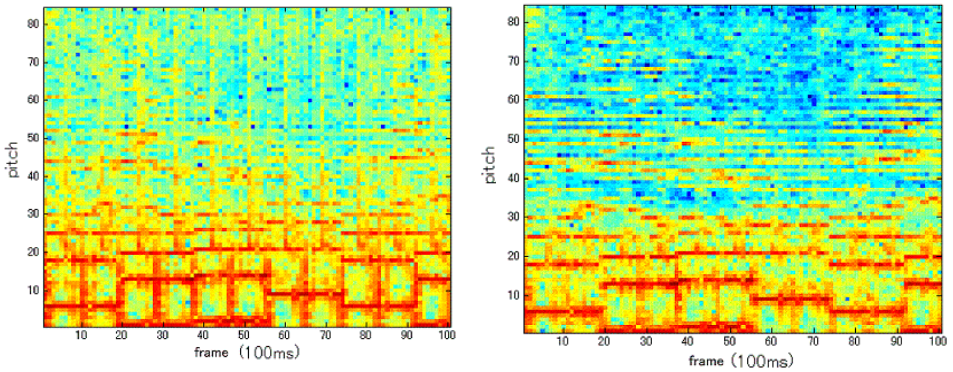

図 1 (左) 元の信号のスペクトログラム, (右) 調波音を強調した信号のスペクトログラム

2 章では本研究の和音認識のフレームワークについて述べる。3 章ではチューニング補正手法について述べる。4 章では実験を行い、和音認識結果の評価を行う。5 章では本稿をまとめ、今後の展望を述べる。

2. 調波音強調と HMM を用いた和音認識のフレームワーク

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、非調波的な成分は和音認識の性能低下の原因となり得る。この問題に対し、信号の調波音の強調を実現する手法として、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] がある。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 (図 1)。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 HMM による和音認識

楽曲では和声内音の省略や非和声音の挿入があるため、単一のフレームでの認識は難しい。和音はある程度の時間持続し、和音間の遷移のしやすさにも偏りがあるため各フレームの和音認識にはそのフレームの特徴ベクトルだけでなく、前後のフレームを用いるのが妥当だと考えられる。現在の時刻の和音は、あまり離れていない時刻の和音の影響が支配的であると考えられため、和音進行は現在の和音の $N-1$ 個前までの和音に依存する N-gram の確率過程として表現でき、特徴ベクトルは隠れ状態である和音から確率的に出力されるという仮定を置く。ここでは和音の移り変わりは直前の和音のみに依存するとして、2-gram でモデル化する。また、和音間の遷移のしやすさは楽曲の調により異なるが、ここでは調を考慮せずすべての調で和音間の遷移確率を一定と近似する。これは調を考慮したモデルよりも粗い近似であるが、転調のある楽曲にも容易に対応できる利点がある。

以上の仮定に基づき、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特徴ベクトル系列を x、求める和音進行系列を c とおくと、和音認識問題はベイズの定理により

$$\operatorname*{argmax}_c p(c|x) = \operatorname*{argmax}_c \frac{p(x|c)p(c)}{p(x)} \tag{4}$$

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、

$$\begin{aligned}&\underset{c}{\operatorname{argmax}}\, p(c|x)\\ &\simeq \underset{c}{\operatorname{argmax}}\, p(x_0|c_0)p(c_0)\prod_{t=1}^{T} p(x_t|c_t)p(c_t|c_{t-1})\end{aligned} \tag{5}$$

となり、和音毎の特徴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、最尤となる和音進行系列 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。

3. チューニングを補正したクロマベクトル

3.1 クロマベクトルの生成

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[5] がある。クロマベクトルは、式 (6) 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。

$$p(k,t) = \sum_{i=0}^{I-1} H(12i+k,t) \tag{6}$$

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

$$f_k = f_{min}2^{k/12} \tag{7}$$

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。

3.2 楽曲間でチューニング一定とみなしたクロマベクトルの問題点

実際の楽曲では楽曲間でのチューニングが一致していない可能性があり、そのような場合にチューニングを一定としてクロマベクトルを生成すると、いくつかの bin にまたがってエネルギーが分配されるなど明瞭なクロマベクトルが得られず、認識率低下の要因となると考えられる。例えば図 3 のように、それぞれ A4 を 440.0Hz と 446.4Hz でチューニングし、ピアノの音色で楽譜 (図 2) を演奏した信号から、チューニングが 440.0Hz で一定であるとしてクロマベクトルを求めると、左の図では和声内音 (ACDF, DFGB, BCEG, ACEF) が明瞭に現れているのに対し、右の図では不明瞭になっていることがわかる。

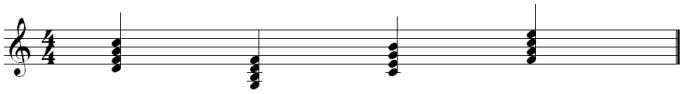

図 2 楽譜 (Dmin7 → G7 → Cmaj7 → Fmaj7 の和音進行)

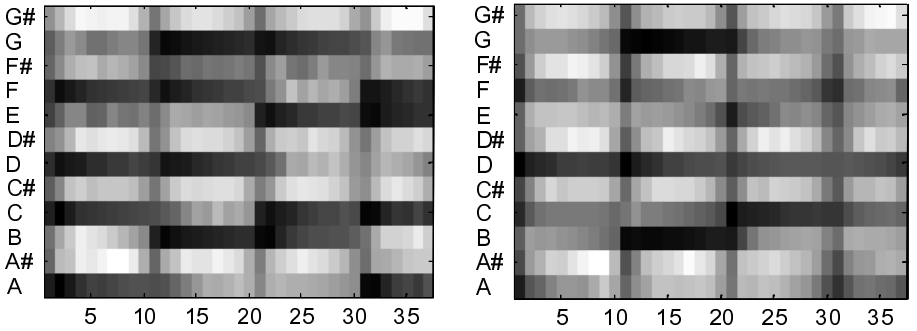

図 3 (左) チューニングがずれていない信号のクロマベクトル、(右) チューニングがずれた信号のクロマベクトル
横軸は 100ms 毎のフレーム

3.3 チューニング補正手法

本節では前節の問題に対処する一つの解決法として以下のようなチューニング補正アルゴリズムを検討する。一つの楽曲中でチューニングが一定であるという仮定の下、正しいチューニングに最も近い中心周波数を持つクロマベクトルのパワーが最大になるということに着目する。理想的にはチューニングが元の楽曲と一致したクロマベクトルが得られることが望ましいが、近似的に、選択できるチューニングの候補を増やし、候補の中からパワー最大のクロマベクトルを選択することにより、理想的なクロマベクトルに近いものを得られると考えられる。これは式 (8) と定式化できる。なお、候補となるクロマベクトルを $p_j(k,t)$ と置いた。チューニングを補正したクロマベクトルは、式 (7) よりスペクトログラ

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

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。$n = 3$ の場合は Mauch ら[8] が行っている。一般に n が大きいほど正確にチューニングを反映できるが、n を大きくするにつれて反映できるピッチの分解能の差が小さくなっていくことや、n に比例して計算時間が増加するため、そこまでの大きな値は実用的でないであろう。そこで、本研究では $n = 1, 3, 5, 7$ の場合について検討を行う。

$$\hat{j} = \underset{j}{\operatorname{argmax}} \sum_{t=1}^{T} \sum_{k=1}^{12} p_j(k, t) \ , j = 1, \cdots, n \tag{8}$$

以上の議論により、本研究では、HPSS により調波音を強調した信号から定 Q フィルタバンクによりスペクトログラムを得て、そこから生成したチューニングを補正したクロマベクトルを特徴量として用い HMM で和音認識を行う。

4. 評 価 実 験

4.1 データセット

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。これらの楽曲は C. Harte[2] による時刻毎の和音ラベル付けがされているため、HMM の学習および認識は状態遷移系列が既知として行うことができる。楽曲はすべてモノラル化し、11025Hz にダウンサンプリングした。

4.2 特徴量の生成と HMM の学習

調波音を強調したクロマベクトルに関する評価は内山らが行っており[6),7]、その有効性は確認されているため、チューニング補正手法の性能の検証を行った。元の楽曲のチューニングが 440Hz で一定であるという仮定の下、90 曲のチューニングを Phase Vocoder を用いて曲ごとに、-50 cent から +50 cent の間でランダムに変化させた。残りの 90 曲で HMM の学習を行い、チューニング変化をさせた楽曲の和音認識を行った。時間周波数分解に用いた定 Q フィルタバンクの Q 値 (周波数/帯域幅) は $Q = 60.0$ であった。Phase Vocoder は Ellis[9] による Matlab プログラムを、HMM の学習および認識には HTK[11] を用いた。

また、チューニングを変化させていない元の楽曲に対し、調波音の強調、チューニング補正を行い、2 分割交差検定により学習、認識を行うことにより従来手法との認識率の比較を行った。これは、内山らが行った実験と同じ条件である。

いずれの場合もクロマベクトルの生成には最低周波数の基準 f_0 を 55Hz とし、最低周波数 f_min から 5 オクターブを用いた。また、和音の種類は表 1 のように major と minor に限り、その他の和音は第 3 音の長・短に基づき分類した。認識率は、フレーム毎の和音ラベルの正解数を、フレーム全体で割ることで算出した。

表 2 チューニングをずらした楽曲 90 曲における候補数と認識率の関係

候補数 (n)	1	3	5	7
認識率	55.9%	67.3%	67.1%	67.6%

表 3 元の楽曲 180 曲における候補数と認識率の関係

候補数 (n)	1	3	5	7
認識率	76.4%	78.2%	78.6%	78.9%

4.3 和音認識結果

チューニングを意図的にずらした楽曲に対し、チューニング補正手法を適用して検出したピッチのずれを図 4 に示す。各 n について、実際のピッチのずれからもっとも近い各中心周波数に検出される傾向にあることがわかる。次に HMM による和音認識結果を表 2 に示す。$n = 1$ の場合の認識率がチューニング補正を行わない内山らの手法での認識率であり、チューニング補正手法により認識率が 11% 程度向上していることがわかる。また、曲ごとの認識結果の分布を図 5 に示す。$n = 1$ の場合と比べ、$n = 3, 5, 7$ の場合の分布が認識率の高い方に寄っていることがわかる。チューニング補正を行った場合・行わなかった場合のいずれにおいても認識率の極端に低い楽曲があるが、これらの楽曲はインド音楽に影響を受けた楽曲など、西洋音楽や調性音楽でないためである。以上の実験結果より、提案法がチューニングのずれが存在する場合の和音認識に有効であることが確認できる。

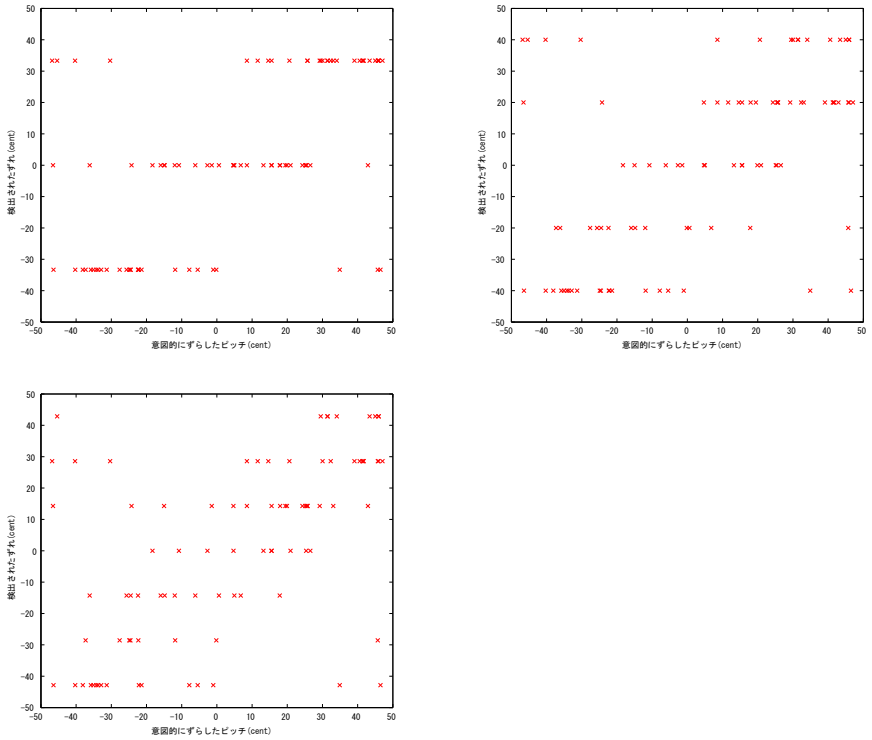

図 4 The Beatles の 90 曲の楽曲ごとの意図的にずらしたピッチ (横軸) と検出されたピッチのずれ (縦軸) の関係 (左上)$n = 3$, (右上)$n = 5$, (左下)$n = 7$

チューニングを意図的にずらさない元の楽曲に対し、同手法を適用した和音認識結果は表 3 であった。$n = 1$ の時の認識率がチューニング補正を行わない内山らの手法での認識率であり、チューニング補正手法により認識率が 2% 程度向上していることがわかる。また、検出されたピッチのずれのヒストグラムは $n = 7$ の場合は図 7 であり、ピッチのずれが 0cent と検出された楽曲数と同じ位、-14.3 cent にも検出された。このことから、対象となる楽曲でチューニングの異なる楽曲が存在し、そのチューニングが補正されたことにより認識率が向上したと考えられる。高々数% の認識率の向上だったため、楽曲ごとの認識結果の分布 (図 6) には大きな差は見られなかった。

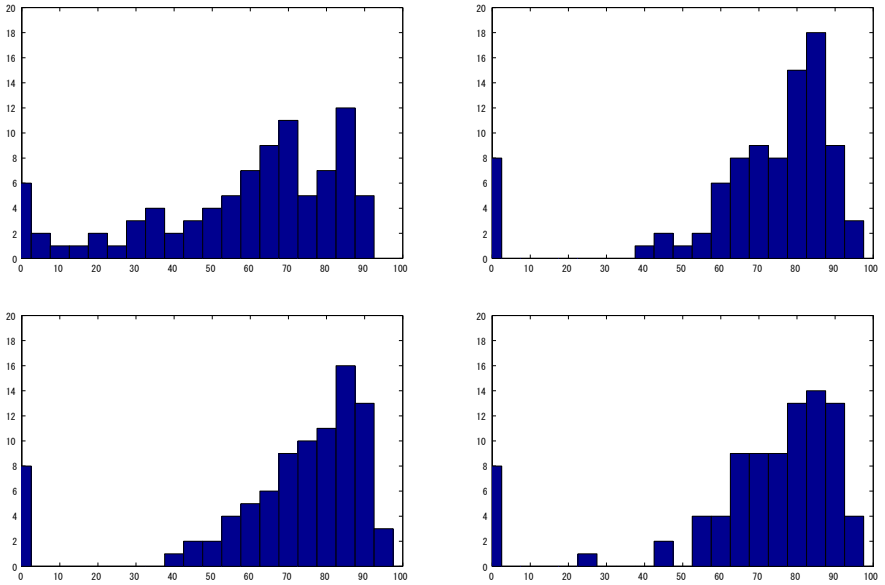

図 5 The Beatles の 90 曲の楽曲ごとの認識率のヒストグラム (横軸) 認識率 (%), (縦軸) 楽曲数 (左上)$n = 1$, (右上)$n = 3$, (左下)$n = 5$, (右下)$n = 7$

5. おわりに

5.1 まとめ

本研究では、音楽音響信号からの自動和音認識の性能向上のために、調波音を強調しチューニング補正をしたクロマベクトルを特徴量として用いることを提案した。まず、調波音／打楽器音分離により調波音を強調した信号から n 個のクロマベクトルの候補を生成した。n 組の中からエネルギー最大となるクロマベクトルを特徴量として用いることで、楽曲ごとのチューニングの違いに対応し、実験により和音認識の性能が向上することを確認した。

5.2 今後の展望

今回の手法では、近似的なチューニング補正を行ったが、より正確にチューニングを推定し補正する手法により、さらに和音認識率の向上が見込まれる。和音の種類に関して、和音認識の実用的な応用を考えると今回実験を行った major, minor の 2 種類 だけで認識を行

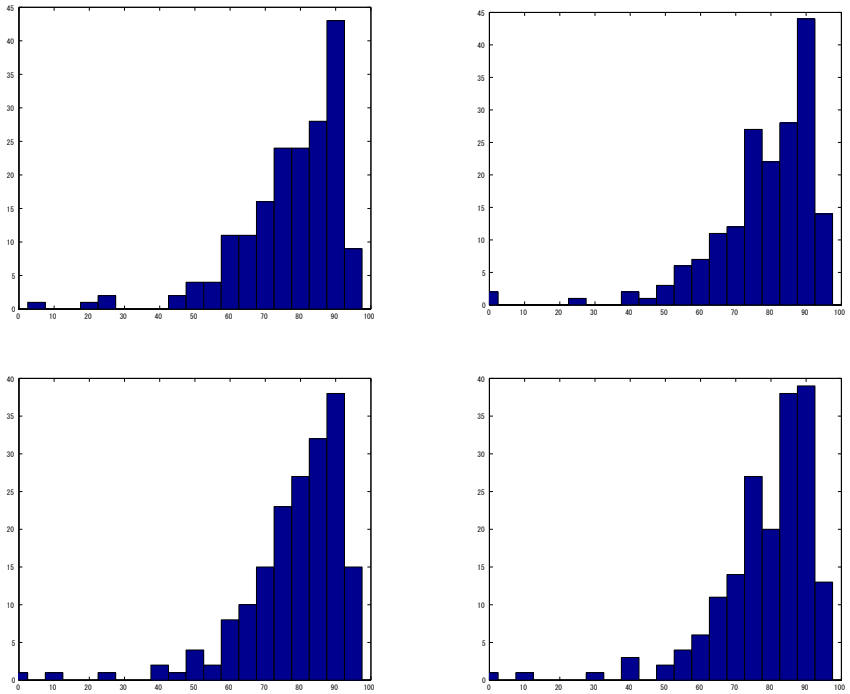

図 6 The Beatles の 180 曲の楽曲ごとの認識率のヒストグラム (横軸) 認識率 (%), (縦軸) 楽曲数 (左上)$n = 1$, (右上)$n = 3$, (左下)$n = 5$, (右下)$n = 7$

うことは現実的でなく、その他のコードラベルも含めて和音認識実験を行うべきであろう。また、クロマベクトルでは半音毎の情報のみを用いているが、さらに低音の情報を利用することで和音の転回形を含めた認識を行うことが考えられる。最後に、現在の手法では調波音を強調した信号から得た特徴量のみを用いているが、和音境界での認識誤りが多い。一般的に、和音はビート位置で変化することに着目し、打楽器音を強調した信号からの特徴量を用いてこの問題を改善することが考えられる。

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。

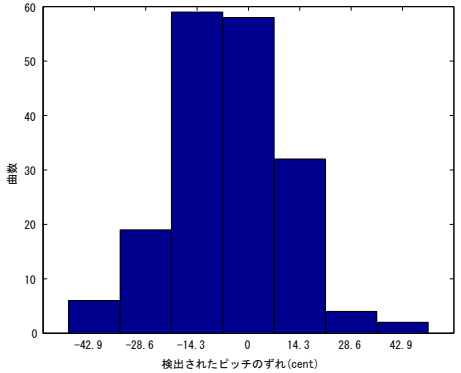

図 7 元の楽曲 180 曲で検出されたピッチのずれ ($n = 7$ の場合)

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) 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.
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) T. Fujishima, "Real-time chord recognition of musical sound: A system using common lisp music," *Proc. ICMC*, pp. 464–467, 1999.
6) 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.
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) M. Mauch *et al.*, "A discrete mixture model for chord labelling," *Proc. ISMIR*, pp. 45–50, 2008.
9) D. Ellis, `http://www.ee.columbia.edu/~dpwe/resourses/matlab/pvoc/`
10) MIREX 2008, `http://www.music-ir.org/mirex/2008/index.php/Main_Page`
11) HTK Hidden Markov Model Toolkit, `http://htk.eng.cam.ac.uk/`