音 楽 情 報 科 学 44－9
（２００２．２．１６）

櫛形フィルタと確率モデルに基づいた音高認識

梶山 晋弘†　　山下 洋一††

†立命館大学大学院理工学研究科情報学システム専攻
††立命館大学理工学部情報学科
〒 525-8577 滋賀県草津市野路東 1-1-1

E-mail:{kaji,yama}@slp.cs.ritsumei.ac.jp

あらまし

音声認識で広く用いられている HMM を用いて音高をモデル化し, 楽曲の音高認識を試みた. 楽器音のスペクトルの特徴と国際基準による 12 平均律音階に基づき,48 個の櫛形フィルタを作成した. そして, これらの櫛形フィルタで楽器音のスペクトルをフィルタリングしたときに得られる 48 個の出力強度を特徴量とした. モデルは MIDI 音源によるフルートとトロンボーンを用いて作成し, それぞれの楽器での孤立単音, ソロによる連続音, 孤立和音, デュオによる連続音に対しての認識を行った.

キーワード：自動採譜, 櫛形フィルタ, 確率モデル, 隠れマルコフモデル, 音高認識

Pitch Recognition of Musical Sounds Based on Comb Filters and a Stochastic Model

Kunihiro Kajiyama†　　Yoichi Yamashita††

†Graduate school of Science and Engineering, Ritsumeikan University
††Dep. of Computer Science, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu-shi, Shiga, 525-8577 Japan

E-mail:{kaji,yama}@slp.cs.ritsumei.ac.jp

Abstract

The pitch of sound is recognized by a HMM technique which is widely used for speech recognition.To extract features of the pitch, 48 comb filters are designed based on 12-temperament scale. The spectrum of musical instrument sound is filtered with 48 comb filters, and a feature vector of the sound consists of the output intensities. The model was trained for the flute and trombone by using MIDI sound data. Performance of the pitch recognition is investigated for the isolated single sound and continuous sound performed by the solo, and for the isolated chord and continuous sound performed by the duet.

Keywords: automatic transcription, comb filter, stochastic model, Hidden Markov Model, pitch recognition of sound

1　はじめに

耳で聞いた音楽を譜面にする技術を採譜という. 採譜は作曲や編曲などの音楽活動をする際に必要な技術である. しかし, 採譜は音楽熟練者だけが持つ特殊な技術であり, 音楽非熟練者にとっては非常に困難なものである. このため, 音楽非熟練者に代わってコンピュータ採譜を行なうシステムが望まれている [1][2]. そこで, 本研究では自動採譜システムを構築することを目指す. 自動採譜を実現するためには音高・楽器の種類・調性の 3 つを認識しなければならない. 現在, 本研究ではこのうち, 音高についての認識を行なっている. これらの処理は未知入力パターンのカテゴリを決定するパターン認識の問題である. パターン認識において, 近年, 確率モデルに基づいた手法が広く用いられている. 確率モデルに基づいた手法は, パターン認識の問題のうち多くの例題が利用可能な場合に有効な手法であり, 音声認識に対して広く用いられている. 現在,MIDI 音源によって多くの楽曲の生成が容易に行なえるため, 本研究においてもこの手法は有効である.

2　手法

本研究では, 音声認識で広く用いられている統計的手法を用いて楽曲の音高認識を試みる.HMM [3][4][5] によって各音高の特徴をモデル化し, 音高のあらゆる連鎖 ($W = w_1$,w_2,...,w_n) を考え, $P(Y|W)$ を最大にする Wを求める. ここで,Y は入力パターン時系列 ($Y = y_1, y_2, ..., y_t$) である. 各音高の特徴の抽出には櫛形フィルタを用いる.HMM は left-to-right 型のトポロジーを使い,3 状態のモデルを作成した.$P(Y|W)$ は,

$$P(Y|W) = P(y_1...y_t|w_1...w_n) = \sum_{i_1<i_2<...<i_{n-1}} P(y_1...y_{i_1}|w_1)P(y_{i_1+1}...y_{i_2}|w_2) ...P(y_{i_{n-1}+1}...y_{i_n}|w_n) \quad (1)$$

である.$P(y_{i_{j-1}+1}...y_{i_j}|w_j)$ は, ある HMM のモデル w_jが特徴ベクトル時系列 Yを出力する確率であり,状態遷移系列 $X = x(0), x(1), x(2), ..., x(T)$ として許される全ての音高系列を考えることによって,

$$P(y_{i_{j-1}+1}...y_{i_j}|w_j) = \sum_X \pi_{x(0)} \Pi_{t=1}^T a_{x(t-1)x(t)} b_{x(t-1)x(t)}(y_t) \quad (2)$$

で計算する. 実際には, これを,

$$P(y_{i_{j-1}+1}...y_{i_j}|w_j) = \max_X [\pi_{x(0)} \Pi_{t=1}^T a_{x(t-1)x(t)} b_{x(t-1)x(t)}(y_t)] \quad (3)$$

で近似する Viterbi アルゴリズムを用いて計算する.a_{ij}は遷移確率,b_{ij}は出力確率を表している. 特徴ベクトルの分布を正規分布 (平均 m_{ij}, 共分散行列Σ_{ij}) で近似し, ベクトル y の出力される確率を

$$b_{ij}(y) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_{ij}|^{\frac{1}{2}}} exp[-\frac{1}{2}(y - m_{ij})^t \Sigma_{ij}^{-1}(y - m_{ij})] \quad (4)$$

で計算する.

3　櫛形フィルタ

楽器音のスペクトルを図 1 に示す. 楽器音のスペクトルは基音と倍音からなる. 基音の周波数は音高によって異なり, 倍音の周波数は基音のほぼ整数倍となる. 本研究では, この性質と国際標準による 12 平均律音階 (110.0Hz〜1661.2Hz までを用いる) に基づき,48 個の櫛形フィルタ $CF_i(i = 1, ..., 48)$ を作成した. 図 2 に C_4(523.25 Hz) の音高に対応する櫛形フィルタ CF_{28}の特性を示す. 作成した櫛形フィルタは, 音高のゆらぎの対応するために, 正規分布の形状を用いて特徴の抽出に巾を持たせた.

4　特徴量の抽出

特徴量抽出の構成を図 3 に示す. 作成した 48 個の櫛形フィルタを用いて, 楽器音のスペクトルをそれぞれフィルタリングする. フィルタ CF_iの出力強度を o_iとし,$O = \{o_1, o_2, ..., o_{48}\}$ を音高

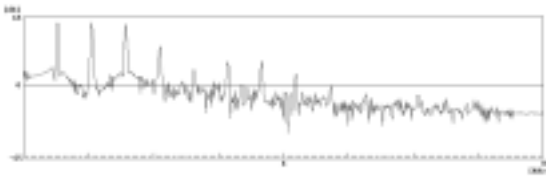

図 1: フルートの (C_4) スペクトル

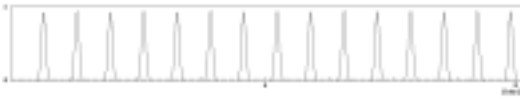

図 2: C_4(523.25Hz) の櫛形フィルタ

の特徴量とする. フィルタリングの処理は周波数領域で楽器音のスペクトルと櫛形フィルタのスペクトルを掛け合わせることによって実現する. 出力強度は楽器音の音高と櫛形フィルタの音高が同じであるとき高い値を示し, 異なるとき低い値を示す. 和音においては含まれる音高と同じ音高である櫛形フィルタの出力強度がどちらも高い値を示す. 出力強度は最大を 1 に正規化した.

5 実験

5.1 データ

5.1.1 モデルデータ

モデルは MIDI 音源によるフルートとトロンボーンを用いて 5 種類の音高モデルセットを作成した. まず, 単一楽器による演奏を認識するためのモデルとして,model-Fa と model-Ta を作成した.model-Fa は楽器音はフルートのみで, 音高は絶対音名で C_3〜C_5の半音階 2 オクターブ,25 音高に無音を加えた計 26 音のモデルとした.model-Ta は楽器音はトロンボーンのみで, 音高は絶対音名で C_2〜C_4の半音階 2 オクターブ,25 音高に無音を加えた計 26 音のモデルとした. 次に, 同一楽器によるデュオ演奏を認識するためのモデルとして,model-Fb と model-Tb を作成した.model-Fb は model-Fa の 25 音高の重複しない組合せ 300 音高に無音を加えた計 301 音のモデルとした.model-Tb は model-Ta の 25 音高の重複しない組合せ 300 音高に無音を加えた計 301 音のモデルとした. 最後に, 異楽器によるデュオ演奏を認識するためのモデルとして,model-FT を作成した.model-FT は model-Fa の 25 音高と model-Ta の 25 音高を全て組合せた 625 音高に無音を加えた計 626 音のモデルとした. サンプリングレートは全て 16KHz とした。

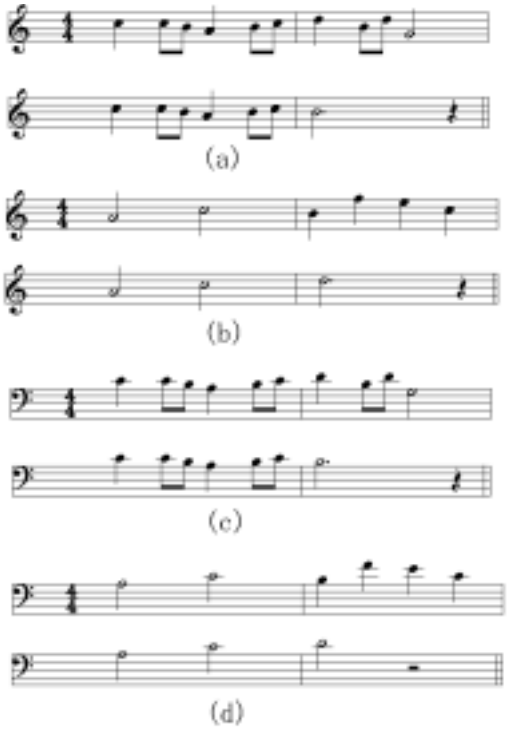

図 4: 評価データ

5.1.2 単音からの和音の合成

モデル作成の際に必要な和音を全て用意しようとすると莫大な量になることが予想される. したがって, 本研究では単音だけを用意し, スペクトル合成によって和音を生成した.

5.1.3 評価データ

モデルごとの評価データを以下に示す. 評価データは全て MIDI 音源を使用し, テンポは 120 とした. 図 4 の譜面 (a) と図 4 の譜面 (c), 図 4 の譜面 (b) と図 4 の譜面 (d) はオクターブ違いの同じメロディとなっている.

- **data-Fa**
  model-Fa 作成に使用した全ての音高の孤

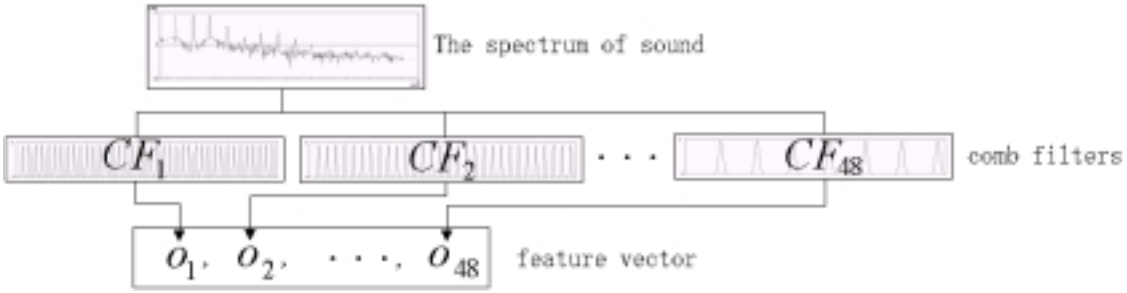

図 3: 特徴量の抽出

立単音を data-Fa1 とする. フルートによる全音階 1 オクターブ (C_4〜C_5) の連続音を data-Fa2 とする. 図 4 の譜面 (a) をフルートで演奏した音を data-Fa2 とする.

- **data-Ta**
  model-Ta 作成に使用した全ての音高の孤立単音を data-Ta1 とする. トロンボーンによる全音階 1 オクターブ (C_3〜C_4) の連続音を data-Ta2 とする. 図 4 の譜面 (c) をトロンボーンで演奏した音を data-Ta3 とする.
- **data-Fb**
  model-Fb 作成に使用した全ての音高の孤立和音を data-Fb1 とする. 図 4 の譜面 (a) と図 4 の譜面 (b) をフルートデュオで演奏した音を data-Fa2 とする.
- **data-Tb**
  model-Tb 作成に使用した全ての音高の孤立和音を data-Tb1 とする. 図 4 の譜面 (c) と図 4 の譜面 (d) をトロンボーンデュオで演奏した音を data-Tb2 とする.
- **data-FT**
  model-FT 作成に使用した全ての音高の孤立和音を data-FT1 とする. 図 4 の譜面 (a) をフルートで演奏し, 図 4 の譜面 (d) をトロンボーンで演奏したフルート・トロンボーンデュオの音を data-FT2 とする.

5.2 認識結果

5.2.1 認識率の算出

認識結果は認識された音高名と認識された音高の開始時間と終了時間を 10ms 単位で示す形で得られる. 認識率を算出するためには正解である音高名と正解である音高の開始時間と終了時間を正確に表した正解データが必要となる. 本研究では, 正解データを表 1 にしたがい人手で作成した. 認識率は, 認識結果と正解データを 10ms 単位で比較し, 音高が同じである割合を求めた.

表 1: 音符ごとの時間長 (テンポ 120)

note	duration
whole note	2 sec
half note	1 sec
quarter note	0.5 sec
8-minute note	0.25 sec
16-minute note	0.125 sec

5.2.2 櫛形フィルタの出力強度による音高認識

出力強度が最大値を示した櫛形フィルタの音高を認識結果としたとき,data-Fa1 に対する認識率は 80.0 %であった. このことから, 出力強度が最大値を示す櫛形フィルタの音高が必ずしも楽器音の音高と一致しないことがわかる.

5.2.3 単一楽器による演奏の認識

単一楽器による演奏の認識率を表 2 に示す. また, フルートソロによるメロディの認識結果の一部を図 8 に示す. 櫛形フィルタの最大値を用いる認識では孤立単音に対しての認識でもオクターブ違いの誤認識が見られたが,48 個の櫛形フィルタの出力を合わせて用いることにより, 孤立単音に対しては 100 %の認識率が得られた. しかし, 全音階 1 オクターブとメロディによる連続音の

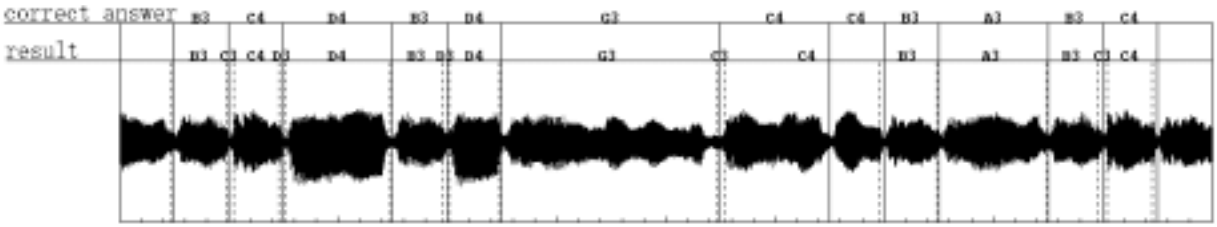

図 5: フルートソロによるメロディの認識結果 (一部拡大)

表 2: 単一楽器による演奏の音高認識率

model	data	rec. rate
model-Fa	data-Fa1	100%
	data-Fa2	93.3%
	data-Fa3	90.3%
model-Ta	data-Ta1	100%
	data-Ta2	92.8%
	data-Ta3	94.0%

認識では若干誤認識が見られた. 誤認識のほとんどは音と音の過渡部に見られた. 過渡部において正解データより認識結果の方が早く音が変化していたり, 正解データの音高より 1 オクターブ低い音高が認識されていた. この過渡部の認識率の低下にはフレーム長の影響があると考えられる. 本システムではフレーム長は 64ms(1024 ポイント) とし, フレーム間隔は 10ms(160 ポイント) としている. このフレーム中に過渡部があると同じフレームに 2 つの音高が含まれる. したがって, この 2 つの音高が含まれるフレームの認識率が低下していると考えられる. さらに, フレームに含まれる次の音高の割合が多くなったときに認識結果が次の音高となるため, 過渡部において正解データより認識結果の方が早く音が変化していると考えられる. また, 楽器の残響音の影響よって音の最後では音符が長くなっていた.

5.2.4 同一楽器によるデュオ演奏の認識

同一楽器によるデュオ演奏の認識率を表 3 に示す. また, フルートデュオによるメロディの認識結果の一部を図 9 に示す. 孤立和音に対する認識では音高は正確に認識された. しかし, メロディに対する認識では, 含まれる和音の片方において, オクターブ違いが見られた. また, フレーム長と残響音の影響はここでも見られた.

表 3: 同一楽器によるデュオ演奏の音高認識率

model	data	rec. rate
model-Fb	data-Fb1	100%
	data-Fb2	73.2%
model-Tb	data-Tb1	100%
	data-Tb2	77.8%

5.2.5 異楽器によるデュオ演奏の認識

異楽器によるデュオ演奏の認識率を表 4 に示す. また, フルート・トロンボーンデュオによるメロディの認識結果の一部を図 10 に示す. 図中の「C3+A2」は楽器がフルートで音高が C_3 である音と楽器がトロンボーンで音高が A_2 である音との和音であることを示す. 孤立和音に対する認識では楽器の種類も含めて正確に認識された. メロディに対する認識では含まれる和音の片方においてのオクターブ違いが多く見られたが, 楽器の種類は正確に認識された. フレーム長と残響音の影響はここでも見られた.

6 おわりに

本稿では, 櫛形フィルタと確率モデルに基づき, 楽器によって演奏された音に対する音高認識

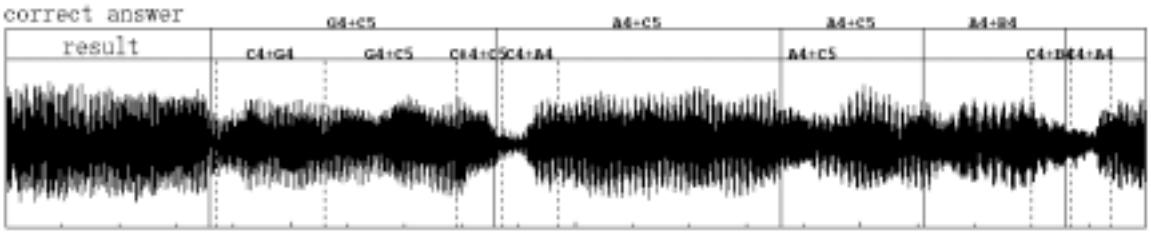

図 6: フルートデュオによるメロディの認識結果 (一部拡大)

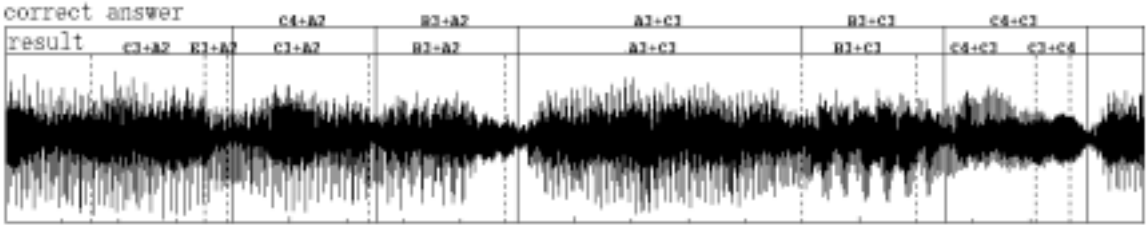

図 7: フルート・トロンボーンデュオによるメロディの認識結果 (一部拡大)

表 4: 異楽器によるデュオ演奏の音高認識率

model	data	rec. rate
model-FT	data-FT1	100%
	data-FT2	68.3%

について述べた. 単一楽器による演奏の認識では音高はほぼ正確に認識されたが一部ではオクターブ違いも見られた. 同一楽器によるデュオ演奏の認識と異楽器によるデュオ演奏の認識では音高はほぼ正確に認識されていたがオクターブ違いが単一楽器による演奏の認識よりも多くの箇所で見られた. 異楽器によるデュオ演奏の認識では楽器の種類が正確に認識された. 全体の認識率を低下させているフレーム長と残響音の影響は時間的にかなり短いため, 実際に譜面を作成する際はそれほど影響しないと思われる. 今後の課題としては, 本システムでは, 同じ音高である音符が複数続いた際にどこで音符が区切られているかを認識することができない. また, 同一楽器によるデュオ演奏では, 和音となっている 1 つ 1 つの音がどちらの楽器から発生した音であるかの識別ができない. したがって本システムの認識結果だけで譜面を作成することが困難であるためこれを解決していく.

参考文献

[1] 長島洋一, 橋本周司, 平賀譲, 平田圭二編. コンピュータと音楽の世界. 共立出版社,1998.

[2] Curtis Roads. コンピュータ音楽. 東京電機大学出版局,2001.

[3] 今井聖. 音声認識. 共立出版社,1995.

[4] 鹿野清宏, 中村哲, 伊勢史郎. 音声・音情報のディジタル信号処理. 昭晃堂,1997.

[5] S.J.Young. P.C.Woodland. W.J.Byrne. HTK User Manual. Entropic Research Laboratory,1993.

[6] 中川聖一. 確率モデルによる音声認識. 電子情報通信学会,1988