情報処理学会研究報告
IPSJ SIG Technical Report

音楽音響信号の低音旋律パターンのクラスタリングと自動ジャンル認識への応用

角尾　衣未留[†1]　George Tzanetakis[†2]
小野　順貴[†1]　嵯峨山　茂樹[†1]

本研究は音楽音響信号中に含まれる小節単位の低音旋律パターンをジャンルごとに複数種類抽出し、ジャンル認識の精度を向上させる事を目的としている。小節単位の低音旋律パターンは例えばロックでは同じ音を同じリズムで演奏されるのに対し、ジャズではウォーキングベースと呼ばれる複雑なパターンであるなど、ジャンル毎に特徴的である。低音旋律パターンのピッチシフトに対する不変性を考慮した k-means クラスタリング法を提案し、ジャンル毎のパターンテンプレートの学習を行い、楽曲に含まれているパターンとテンプレートの距離に基づく特徴量ベクトルを算出することによって、ジャンル認識を行いその有効性を検証する。

Bass-Line Pattern Clustering from Audio Signals and Its Application to Automatic Genre Classification

Emiru Tsunoo,[†1] George Tzanetakis,[†2]
Nobutaka Ono[†1] and Shigeki Sagayama[†1]

This paper discusses a new approach for clustering musical bass-line patterns representing particular genres and its application to audio genre classification. Many musical genres are characterized by not only timbral information but also representative bar-long bass-line patterns. For instance, while a bass-line in rock music is constant pitch and a uniform rhythm, in jazz music there are many characteristic movements such as walking bass. We propose a representative bass-line pattern template extraction method based on k-means clustering handling a pitch-shift problem. After extracting the templates for each genre, a feature vector is calculated and used for automatic genre classification.

†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
†2 ヴィクトリア大学コンピュータサイエンス科
Computer Science Department, University of Victoria

1. はじめに

近年のインターネット配信やポータブルプレイヤに利用される音楽ファイル増加に伴い、音楽情報検索の研究が盛んに行われるようになった。関連する研究の中でも、特に音楽ジャンル分類によるラベル付けは非常に重要なタスクとして長い間取り組まれている。このタスクにおいて音色情報やピッチ情報、リズム情報などが用いられてきた[1)]。その中で最も有効なものは MFCC などスペクトログラムから計算される音色特徴量であるとされ、またピッチ情報についても、特にベース音特徴量を用いた研究[2),3)] が多くなされてきている。しかしこれらの特徴量は、楽曲全体の性質を捉えるための統計情報として扱われており、楽曲中に部分的に現れる特徴的なパターンや、それらの繰り返し構造などが効率的に表現されていなかった。

これに対し我々は、楽曲が含んでいるより詳細な時間構造をジャンル認識に応用することを目的に、小節単位の打楽器パターンを利用したジャンル認識手法を提案してきた[4)]。従来では困難であった小節境界未知の複数リズムパターン抽出を反復的なクラスタリングにより可能にし、抽出されたリズムパターンを従来の音色情報と組み合わせる事で性能向上を実現した。このような小節単位のパターンは打楽器のみでなく、低音旋律においても特徴的に現れ、ジャンル認識に対して有力な特徴になり得ると考えられる。

本報告では、入力音響信号からジャンル毎に小節単位の低音旋律パターンを抽出する手法を提案し、それらに基づく特徴量を抽出することによるジャンル認識について議論する。まず、音響信号中の調波音を強調し小節への分割推定を行う。分割された小節単位の低音旋律を k-means クラスタリングのアルゴリズムを応用した手法によりクラスタリングし、ジャンルを代表する複数パターンを学習する。最後に、提案された低音旋律パターン情報がジャンル認識に有効である事を実験によって検証する。

2. 小節単位の低音旋律パターン学習

2.1 問題設定

小節単位の低音旋律パターンはジャンル毎に、少なくとも一部のジャンルにおいては特徴的であり、それらの情報はジャンル認識を行う上で非常に有利に働くと考えられる。そこで、ジャンル毎に共通のそのようなパターンを学習データから抽出することができれば、ジャンル推定の手がかりの一つとなり得る。

音楽音響信号から小節単位に切り出された低音旋律を得ることができればそれらを複数のクラスタにクラスタリングすることにより共通単位パターンが学習できると考えられる。しかし、同じパターンが演奏される場合でも楽曲の調や和声 (コード) によって実際に演奏

される音は異なる。例えばコードのルート音を一定のリズムで演奏されるパターンがある時、そのパターンはコードが変化する度に音高方向にシフトされて演奏されるなどである。そのため、ピッチシフトに対して不変であるようなパターンを抽出する必要がある。

さらに、一般の音楽音響信号を対象とする場合には、低音旋律だけが切り出されて観測されているわけではない。通常、楽曲は低音旋律以外のパートも含み、特に打楽器音が含まれる場合には単純なフィルタリングでは除くことができず、低音旋律解析の妨げになることが多い。また、小節単位の低音旋律パターンを抽出するためには小節への分割の推定も必要である。

以上の問題は以下の三点にまとめられる。

(1) 楽音は調波音と打楽器音の混合であること
(2) 小節の分割が未知であること
(3) ピッチシフトを考慮して複数の低音旋律パターンを推定する必要があること

2.2 打楽器音と調波音の分離

一般に楽曲は旋律楽器のみでなく打楽器音も同時に演奏される事が多い (問題 1)。そのため、低音旋律パターンを推定する際にこれらを分離し調波音を強調することによってより正確なパターン抽出を行うことができると考えられる。

打楽器音と調波音を分離する手法の一つとして宮本らの手法[5] を利用することが考えられる。図 1 の左上に典型的なスペクトログラムの一部を示す。これを見ても分かるように、打楽器音と調波音それぞれが異なる音色の特徴を一般に持つ：調波音は特定の周波数で比較的長時間演奏されるためスペクトログラム上では横に滑らか、逆に打楽器音は広い周波数帯域にエネルギーが広がり時間的に瞬時であるので縦に滑らかとなる。このスペクトログラムの滑らかさの異方性を利用し、MAP 推定の枠組で各成分を推定する。この手法を用いて先のスペクトログラムに適用し、調波音と打楽器音を分離したものがそれぞれ図 1 の右上と左下である。このように分離強調された旋律楽器音のみを用いる事で、低音旋律パターンを抽出する。

2.3 打楽器パターンクラスタリングによる小節へのセグメンテーション推定

問題 2 である小節の推定を行うにあたり、一小節にいくつビートが存在するかが不明である場合、ビートを捉える方法よりパターンマッチングの方が効果的であると考えられる。特に打楽器音は小節単位で単純に繰り返すものが多いため、我々が提案する打楽器音からの小節境界推定手法[7] を用いる事が出来ると考えられる。

この手法では前節で用いた音源分離手法を前処理として打楽器音のみが用いられ、小節単位の打楽器パターンとその小節への最適なセグメンテーションが同時に推定される。初期の複数の打楽器パターンを基に動的計画法でアラインメントを計算し、k-means クラスタリングに基づいてテンプレートパターンを平均計算により更新する。この二つのステップを繰り返すことにより収束後には、最適なテンプレートとセグメンテーションが得られ、ここで

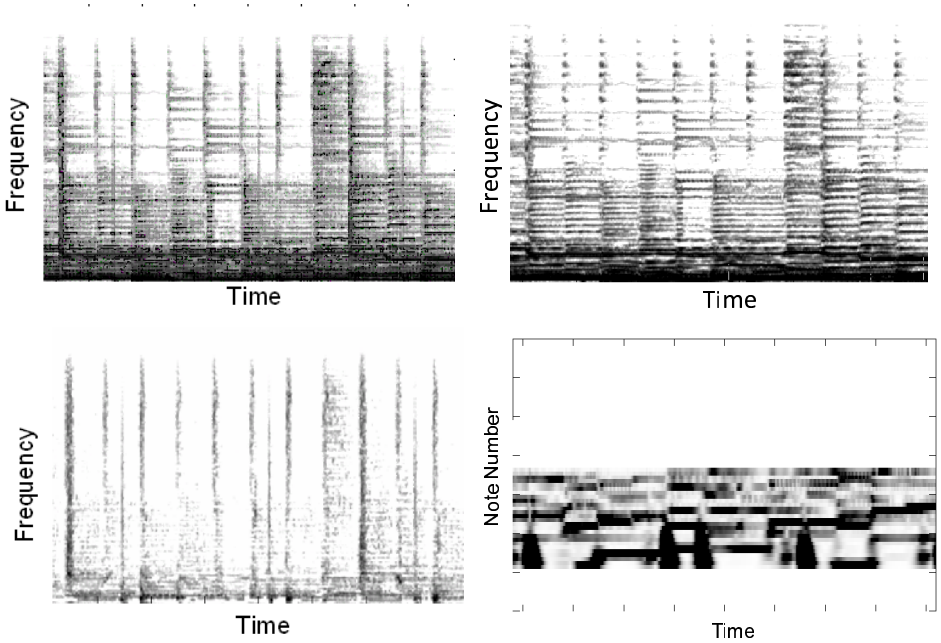

図 1 ポピュラー音楽 (RWC-MDB-G-2001 No. 6[6]) のスペクトログラム (左上) と調波音が強調 されたスペクトログ ラム (右上)、打楽器音が強調されたスペクトログラム (左下)。右下は調波音のみの信号から Wavelet 変換し、ローパスフィルタ処理を行った 対数周波数のスペクトログラム

Fig. 1 The original spectrogram (upper left), the harmonics-emphasized spectrogram (upper right) and the percussion-emphasized spectrogram (lower left) of a popular music piece (RWC-MDB-G-2001 No.6[6]). The low-pass filtered logarithmic spectrogram calculated using wavelet transform from harmonics-emphasized signal is shown in lower right figure.

得られるセグメンテーション情報を用いる事で小節境界を与える事ができる。

2.4 低音旋律パターンテンプレートの反復的更新

最後の問題、問題 3 はピッチシフト問題であり、仮にその問題が存在しなければ単純に k-means クラスタリングを用いる事で複数の代表パターンを抽出する事は容易であると考えられる。ここでは k-means クラスタリングの枠組でピッチシフト不変となるような距離尺度を定義し、新しいクラスタリング手法を提案する。アルゴリズムの流れを図 2 に示す。

一つ留意しておきたいのは、ピッチシフトを考慮する場合には音高ごとのエネルギーを示したスペクトログラム (通常のスペクトログラムを周波数方向に対数的にしたもの) を用い

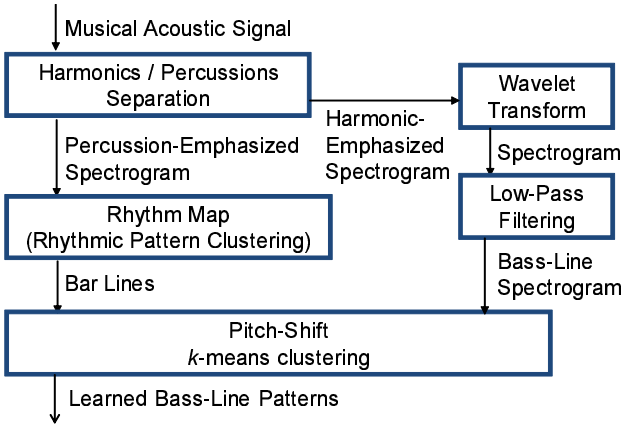

図 2 提案システムのダイアグラム
Fig. 2 The flow diagram of the system.

る事により、容易に処理を行う事が可能となることである。このようなスペクトログラムはウェーブレット変換を用いる事で得られる。図 1 の右下にウェーブレット変換を用いた対数周波数のスペクトログラムを示す。この図では音源分離後にも関わらずバスドラムのエネルギーが強く残っている。これはバスドラムが時に長く響き、スペクトログラム上で時間軸方向に滑らかになるという調波音の特徴を含むことがあるからである。しかし、このエネルギーはテンプレートを更新する際に比較的小さくなるため、低音解析にはそこまで影響を与えないと考えられる。

ピッチシフトを考慮したパターン間距離を計算する一つの方法として、列に半音の音番号を、行に時刻を定義することにより、低音旋律パターンを行列としてモデル化することができると考えられる。もし、入力楽曲に M 個の小節がある場合、m 番目のパターン行列は $I \times N$ 行列 X_m と表せる。ただし、N は低音帯域に存在する半音数、I は小節の時間解像度 (一小節を I 等分) である。一方、k 番目のテンプレートは $I \times 2N$ 行列 B_k と表せる。なぜならば、入力楽曲は N 半音のピッチシフトの可能性があるため、少なくとも N 列のマージンが用意されている必要があるからである。

二つの低音旋律パターン行列間の距離を

$$d(X,Y) = \sum_{i=1}^{I}\sum_{n=1}^{N}(x_{i,n} - y_{i,n})^2 \tag{1}$$

とする。ただし $x_{i,n}$ と $y_{i,n}$ は行列 X、Y の (i,j) 要素である。この場合、ピッチシフトを考慮したテンプレート B_k と小節パターン X_m 間の距離は

$$D(X_m, B_k) = \min_{1\le n\le N} d(X_m, B_{k,n}) \tag{2}$$

と表す事によりピッチシフト不変となる。ここで、$B_{k,n}$ は B_k の n 列から $(n+N-1)$ 列の $I \times N$ の部分行列である。

テンプレート更新に関しては k-means クラスタリングと同様、それぞれの入力低音旋律パターン X_m に関して、距離 (式 2) を最小とするクラス $\hat{k}$ とピッチシフト量 $\hat{n}$ にしたがって平均化を行う。これらは

$$\hat{k}_m = \operatorname*{argmin}_{1\le k\le K} D(X_m, B_k) \tag{3}$$

$$\hat{n}_m = \operatorname*{argmin}_{1\le n\le N} d(X_m, B_{\hat{k},n}) \tag{4}$$

と表すことができる。平均化は E を $N \times 2N$ の拡張行列、S を $2N \times 2N$ のピッチシフト行列として

$$B'_k = \frac{\sum_{m\in\{m|k=\hat{k}_m\}} S^{\hat{n}_m} E X_m}{\sum_{m\in\{m|k=\hat{k}_m\}} 1} \tag{5}$$

$$E = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \tag{6}$$

$$S = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix} \tag{7}$$

となる。

この更新規則はユークリッド距離が使われている限り次の式を満たすため反復更新の収束性が保証される。

$$\begin{aligned} \sum_{m\in\{m|k=\hat{k}_m\}} D(X_m, B'_k) &\le \sum_{m\in\{m|k=\hat{k}_m\}} d(X_m, B'_{k,\hat{n}}) \\ &\le \sum_{m\in\{m|k=\hat{k}_m\}} D(X_m, B_{k,\hat{n}}) \end{aligned} \tag{8}$$

収束後に K 個の低音旋律パターンテンプレート B_k $(k = 1, \ldots, K)$ が得られる。

3. 特徴量抽出

上のようなアルゴリズムを用いて学習されたジャンル内で共通の低音旋律パターンがジャンル間で有意に異なる事を期待した場合、理想的にはジャンル未知の楽曲に対し、いずれのパターンが含まれているかを距離計算 (式 (2)) しクラス属性を解析する事によってジャンルの認識が可能である。しかし、実際にはジャンルによっては明確にパターンが異なる事は稀であったり、ある時刻ではそのジャンルのパターンは含まれているが、違う時刻には他のジャンルのパターンも含まれているなどの場合が考えられる。そのため、そのような事象による認識誤差を減らすため、テンプレートとの距離に基づく特徴量ベクトルを利用し、統計的識別手法を用いて音楽ジャンルの認識を行う方法が考えられる。この目的のためにはサポートベクトルマシーン等の教師あり学習分類器が広く利用されている。

特徴量抽出における一つの方法として、楽曲の各小節毎に全ての学習されたテンプレートと距離計算し、それを楽曲で平均化して、ベクトルとして表現することが考えられる。仮に楽曲が二つ以上のテンプレートの低音旋律を持っている場合でもそれぞれのテンプレートとの距離はジャンルによって特徴があると考えられるからである。例えばブルースの楽曲に含まれる低音旋律パターンはたとえブルースのテンプレートとの距離が他のテンプレートとの距離と比べて最小とならなくとも十分に小さくなるということである。

具体的に G 種のジャンルからそれぞれ K 個のテンプレートを学習した場合、テンプレートは合計 $K \cdot G$ 個学習されたことになり、M 小節の楽曲における平均化された各テンプレートとの距離は

$$d_l = \frac{1}{M} \sum_{m=1}^{M} D(X_m, B_l) \tag{9}$$

で、l は $1 \leq l \leq KG$ であるテンプレート番号である。特徴量ベクトル $\boldsymbol{x}$ は

$$\boldsymbol{x} = (d_1, d_2, \ldots, d_{KG})^T \tag{10}$$

と表すことができる。この特徴量ベクトルを教師あり学習分類に用いることによりジャンルを認識する。

4. アルゴリズムの手順のまとめ

ここまで述べたアルゴリズムをまとめると以下の手順となる。

(1) 前処理
  (a) 調波音・打楽器音の分離
  (b) ローパスフィルタ処理
  (c) 打楽器パターンを用いて小節境界線の推定
(2) 低音旋律パターンクラスタリング
  (a) 初期の (ランダム値の) テンプレートを用意する
  (b) 式 (2) ピッチシフトしながらテンプレートをマッチング
  (c) テンプレートを式 (6) に従い更新
  (d) ステップ (b) と (c) を収束するまで反復更新
(3) ジャンル分類
  (a) 各テンプレートとの距離を計算し (式 (10)) 楽曲を特徴づける特徴ベクトルとして利用
  (b) 機械学習技術を用いてジャンルを分類する

5. 評 価 実 験

5.1 データセット

GTZAN データセット[1] を用いて評価実験を行った。このデータセットは Blues, Classical, Country, Disco,Hiphop, Jazz, Metal, Pop, Reggae,Rock の 10 ジャンルでそれぞれ 100 曲であり、22.05kHz、1ch 信号にダウンサンプリングしたものを用いた。全ての楽曲に対し音源分離手法を用いて調波音を分離強調し、ウェーブレット変換を行った。そしてローパスフィルタ処理を行い 82.4Hz (E1) から 330.0Hz (E3) の帯域の低音旋律を抽出した。82.4Hz 以下の帯域を用いなかったのはその帯域でのバスドラムのエネルギーの影響が大きかったためである。

5.2 テンプレートの学習と特徴量抽出

まずは前述のアルゴリズムに基づいて各ジャンルで共通の低音旋律パターンを学習した。提案アルゴリズムは音楽情報検索に特化したオープンソースソフトウェアである音響処理フレームワーク、*Marsyas*[⋆1] を用いて実装した[8]。テンプレート学習を一般化するため、各ジャンルのデータを半分に分けた 50 曲でジャンルごとに 2 組のテンプレートを学習した。実験では十分収束が見られたため反復回数を 15 回に固定した。また、各パターンの時間解像度は 16 等分で固定した。

学習されたジャズ、ブルース、ロック、ヒップホップのテンプレートの例を図 3 に示す。ジャズのパターンには動きが見られ、ロックのパターンは水平な直線が見られる。ブルースのテンプレートには典型的なスイングのリズムが見られ、ヒップホップには強いドラム音は演奏されても低音旋律はメロディックに演奏されない事からテンプレートでは低音旋律は疎らであった。このように見ると、ジャンルごとの特徴が各テンプレートに学習された事が言えるであろう。

2 つのセットのテンプレートを学習した次に、特徴量ベクトル (式 (10)) を求めた。二つに分けたうちのデータ 1 から学習されたテンプレートを用いてもう半分のデータ 2 で特徴

⋆1 http://marsyas.sness.net/

表 1 低音旋律パターン特徴量のみの場合のジャンル認識率 (上、ベースラインは 10%) と従来の音色特徴量と組み合わせた場合の認識率 (下)
Table 1 Genre classification accuracy using only bass-line pattern features and ones merged with timbral features

Features	data 1	data 2
Baseline (random classifier)	10.0%	10.0%
Only bass-line (400 dim.)	42.0%	44.8%
Existing (timbre, 68 dim.)	72.4%	72.4%
Merged (468 dim.)	74.4%	76.0%

量を抽出、また、データ 1 に関する特徴量は逆である。

5.3 特徴量による分類とその結果

特徴量空間における分類器の学習については、マシンラーニングのツールキット、“Weka”を用いた[9]。識別器として線形 SVM を用い、10 フォールドの交差確認を行った。つまり、それぞれのセットのデータに対してラベル付けされたデータが 10 サブセットに分けられ、巡回的に 1 サブセットがテスト用、残りの 9 サブセットのデータが訓練用に用いられる評価方法である。

各ジャンルから学習するテンプレート数を変化させたときの認識率を示したものが図 4 である。特徴量ベクトルの次元は一つのジャンルから学習されるテンプレート数が K の時 $10K$ となる。これから見てとれるように、提案特徴量は十分にジャンルの情報を表していると言え、学習するテンプレート数を増やすにつれ認識率は向上した。

次に、従来のジャンル認識システムとの統合した場合の実験を行った。MIREX 2008[10] のジャンル認識のタスクで最も高い認識率であったシステムでは音色に関する特徴量、つまり MFCC 等のスペクトログラムに関する情報の統計量を用いたもの (68 次元) を用いていた。この特徴量と提案する低音旋律パターン特徴量とを合わせた特徴量での認識結果を図 5 に示す。テンプレート数 $K = 40$ の時に最も認識率が高く、それ以上次元を増やした際「次元の呪い」と呼ばれるオーバーフィッティングが起こったため、認識率の低下が見られた。特に $K = 40$ とした時の詳細な認識率は表 1 に示す。これらの結果は従来の音色のみに依ったジャンル認識より高い認識率を示しており、提案した低音旋律パターン情報の有効性を確認した。

6. おわりに

本研究では、音楽音響信号からのジャンル認識を目的とした特徴量抽出として、ジャンル内で共通の小節単位の低音旋律パターンを抽出するアルゴリズムを提案した。まず、調波音・打楽器音分離手法で強調された調波音をローパスフィルタ処理を施し低音旋律を抽出した。そして、打楽器パターンから小節境界を推定した。ピッチシフトに対してパターンが不変となるような新しい k-means クラスタリング手法を提案し、ジャンルを代表する低音旋律テンプレートの学習方法を示した。そしてジャンル認識のため、学習されたテンプレートとの距離に基づく特徴量ベクトルを定義し、実験によって様々な楽曲のデータに対して適用する事により既存のジャンル認識システムを上回る認識率によって低音旋律パターン情報の有効性を確認した。

今後の課題としては、時間解像度等のパラメータを変えた実験や、前後の小節情報も利用した特徴量の考案が考えられるであろう。また、既に提案済の打楽器パターン情報との統合的なシステム構築を行い、さらに認識率を向上させることも考えられる。

謝　辞

本研究の一部は CrestMuse Project の支援を受けて行われた。

参 考 文 献

1) Tzanetakis, G. *et al.*, “Musical genre classification of audio signals,” *IEEE Transaction on Speech and Audio Processing*, pp. 293–302, 2002.
2) Mckay, C. *et al.*, “Automatic genre classification using large high level musical feature sets,” in *Proc. of ISMIR2004*, pp. 525–530, 2004.
3) Tsuchihashi, Y. *et al.*, “Using bass-line features for content-based MIR,” in *Proc. of ISMIR2008*, pp. 620–625, 2008.
4) Tsunoo, E. *et al.*, “Audio Genre Classification Using Percussive Pattern Clustering Combined with Timbral Features,” *Accepted for ICME2009*, 2009.
5) Ono, N. *et al.*, “A real-time equalizer of harmonic and percussive componets in music signals,”in *Proc. of ISMIR2008*, pp. 139–144, 2008.
6) Goto, M., Hashiguchi, H., Nihsimura, T., and Oka, R., “Rwc music database: Music genre database and musical instrument sound database,” in *Proc. of ISMIR2003*, pp. 229–230, 2003.
7) Tsunoo, E. *et al.*, “Rhythm map: extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals,” in *Proc. of ICASSP2009*, pp. 185–188, 2009.
8) Tzanetakis, G., “Marsyas-0.2: A case study in implementing music information retrieval system,” chapter2, pp.31–49, Idea Group Refernece, 2007, Shen, Shepherd, Cui, Liu (eds).
9) Witten, I. *et al.*, “Data mining: Practical machine learning tools and techniques,” *Morgan Kaufmann*, 2005.
10) Mirex 2008, http://www.music-ir.org/mirex/2008/

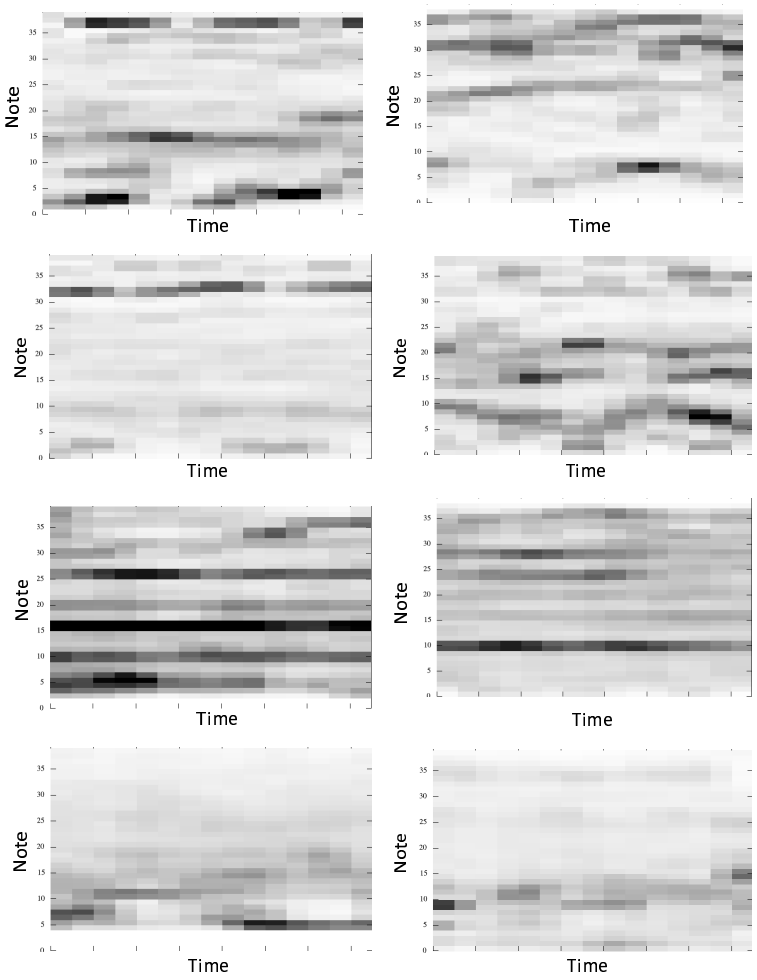

図 3 ジャズ、ブルース、ロック、ヒップホップにおける学習された低音旋律テンプレート (降順)。ジャズやブルースには低音旋律に多くの動きが見られるのに対し、ロックは音高が変化せず、ヒップホップはまばらなパターンとなった。

Fig. 3 Two examples of learned bass-line templates from jazz, blues, rock, and hiphop, in descending order. While a lot of movements of bass notes are shown in jazz and blues, bass notes in rock are unchanged, and hiphop bass-line templates are quite sparse.

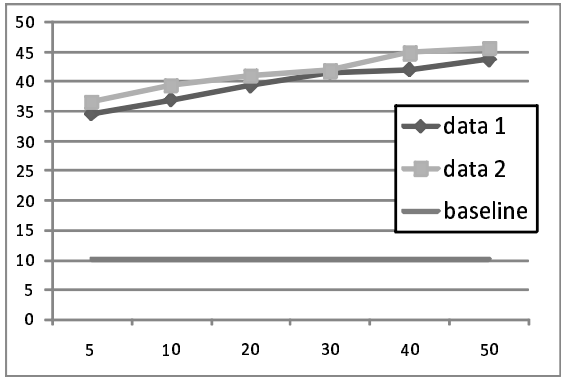

図 4 一つのジャンルから学習されたテンプレート数と低音旋律特徴量のみを用いたジャンル認識率。ランダム分類器によるベースラインは 10%。

Fig. 4 Classification accuracy using only bass-line features for each number of templates learned from one genre. The baseline accuracy of random classifier was 10.0%

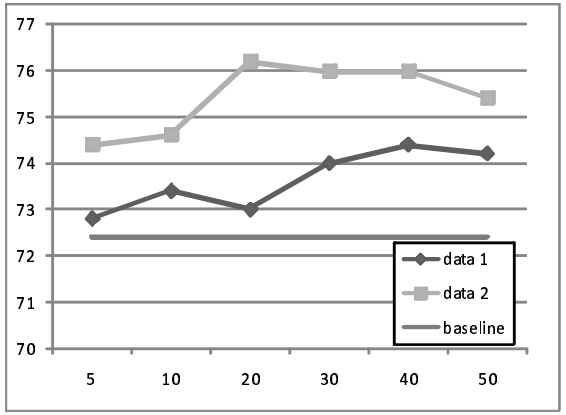

図 5 一つのジャンルから学習されたテンプレート数と低音旋律と音色の特徴量によるジャンル認識率。音色特徴量のみによる認識率は 72.4%。

Fig. 5 Classification accuracy using both bass-line and timbral features for each number of templates learned from one genre. The baseline accuracy of existing features was 72.4%.