Vol.2014-MUS-102 No.25
2014/2/24

Query by Phrase: 半教師あり非負値行列因子分解を用いた音楽信号中のフレーズ検出

増田 太郎[1,a)]　吉井 和佳[2]　後藤 真孝[2]　森島 繁生[1]

概要：本稿では，楽曲群の中から，ユーザがクエリとして与えたフレーズが演奏されている時刻を検出する新たな音楽情報検索手法「Query by Phrase (QBP)」を提案する．本研究において，「フレーズ」とは，何らかの楽器（通常は単一の楽器）による短時間の演奏（音響信号）を指すものとする．QBP では，様々な楽器音が重なり合って構成されている音楽音響信号から所望のクエリを検出する必要がある．この問題を解決するため，本稿では，クエリと楽曲の一部の構成要素との距離を計算することができる手法を提案する．まず，ガンマ過程非負値行列因子分解（GaP-NMF）を用いて，クエリのスペクトログラムを適切な個数の基底スペクトルと対応するアクティベーションとの積に分解する．同様に，楽曲のスペクトログラムも GaP-NMF を用いて分解を行う．このとき，楽曲中にクエリが含まれているという仮定のもと，基底スペクトルの一部にクエリから得られた基底スペクトルをそのまま再利用する（半教師あり GaP-NMF）．こうすることで，クエリにおける基底スペクトルのアクティベーションと楽曲におけるアクティベーションとの類似度を計算することができる．実験により，提案手法は従来の単純なマッチング手法より優れた QBP 精度を達成できることを確認した．

1. はじめに

音響信号をクエリとして楽曲を検索する音楽情報検索 (Music Information Retrieval, MIR) システムが長年にわたり研究されてきた．例えば，類似度に基づく検索システムでは，クエリの音響特徴量に類似した音響特徴量をもつ楽曲を検索することができる [1, 2]．また，音響信号のフィンガープリントに基づく検索システムでは，音響圧縮フォーマットや雑音による音質劣化に頑健な音響特徴量を使用することで，クエリに厳密に一致する楽曲を検索することができる [3, 4]．一方，Query-by-Humming (QBH) システムでは，ユーザの歌唱あるいはハミングによって入力された主旋律を含む楽曲を検索することができる．しかし，楽曲データベース中に MIDI ファイルなどの楽譜情報を登録しておかなければならなかった [5, 6]．この制限を取り払うため，データベース中の楽曲から自動で主旋律を抽出するという研究もなされてきた [7]．

本稿では，混合音である楽曲群の中から，クエリのフレーズに類似するフレーズが登場する時刻を検出する問題に取り組む．我々は，この新しい検索形態を *Query by Phrase* (QBP) と呼ぶことにする．ここで，「フレーズ」とは，数秒間程度の楽器演奏（通常は単一の楽器による）のことである．QBH とは異なり，検索したいフレーズは主旋律である必要はなく，伴奏であってもよい．QBP が実現できれば，一般のユーザにとっては，楽曲名を知らない/忘れてしまった場合でも，その楽曲の特徴的なフレーズを演奏するだけで，直感的に楽曲検索ができる．一方，音楽の専門家にとっては，クエリとして与えたフレーズが，既存の楽曲中でどのように編曲されているかを学ぶのに役立てることができる．

QBP では，様々な楽器音が重なり合って構成されている音楽音響信号から所望のクエリを検出する必要がある．すなわち，クエリと楽曲の一部との距離を適切に計算する必要がある．単純には，自動採譜技術 [8–10] を用いて音楽音響信号を楽譜に変換し，そのうえで音符記号同士の距離を計算する方法が考えられる．しかし，市販 CD のような複雑な混合音に対する自動採譜は非常に困難であり，現時点では現実的ではない．一方，従来の音響特徴量に基づく距離計算手法では，楽曲中に含まれるフレーズに対して他の楽器音が多数重畳することで，音響特徴量が歪んでしまうため，適切な距離が計算できない．

本稿では，クエリと楽曲の “一部の” 構成要素との距離を

[1] 早稲田大学
Waseda University
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
[a)] masutaro@suou.waseda.jp

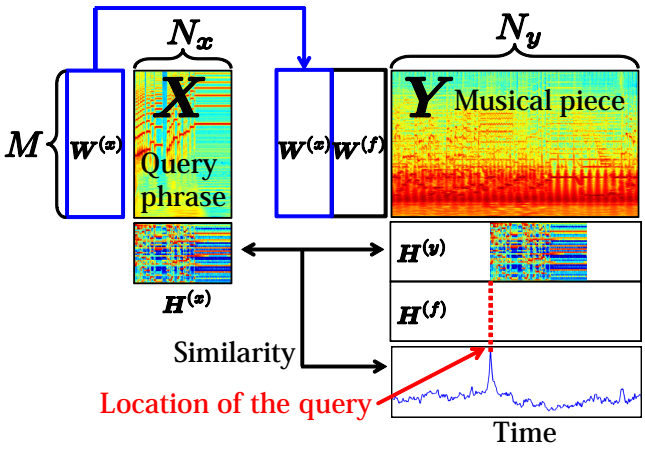

図 1 提案する Query-by-Phrase 検索手法の概要.

計算することで，クエリのフレーズの登場時刻を検出することができる QBP 手法を提案する．本手法は，人間は全楽器パートの完全な採譜ができなくても，楽曲中にフレーズが含まれているかどうかであれば判定できることが多いという現象に着想を得ている．具体的には，まず，ガンマ過程非負値行列因子分解（GaP-NMF）を用いて，クエリのスペクトログラムを適切な個数の基底スペクトルと対応するアクティベーションとの積に分解する．同様に，楽曲のスペクトログラムも GaP-NMF を用いて分解する．このとき，楽曲中にクエリが含まれているという仮定のもと，基底スペクトルの一部にクエリから得られた基底スペクトルをそのまま再利用する（半教師あり GaP-NMF）．こうすることで，クエリにおける基底スペクトルのアクティベーションと楽曲におけるアクティベーションとの類似度を計算することができる．

本稿の残りの章は次のように構成されている．2 章では，提案手法である半教師あり NMF を利用したフレーズ検出手法について説明する．3 章では，提案法との比較のために，従来のマッチング手法を 2 つ紹介する．4 章では，提案手法および従来手法の QBP 精度を評価する実験について述べる．最後に 5 章で，結論と今後の課題について述べる．

2. フレーズ位置検出手法

本章では，ノンパラメトリックベイズ NMF に基づく提案手法である，フレーズ位置検出手法について説明する．

2.1 概要

本手法の目的は，混合音である楽曲群からフレーズの演奏開始点を検出することである．提案手法の概要を図 1 に示す．ここで，$\boldsymbol{X} \in \mathbb{R}^{M \times N_x}$ と $\boldsymbol{Y} \in \mathbb{R}^{M \times N_y}$ はそれぞれクエリおよび検索対象から得た非負値のパワースペクトログラムを表す．本手法は 3 つの手順によって構成される．

まず，クエリ $\boldsymbol{X}$ を NMF を適用することにより，基底スペクトルの組 $\boldsymbol{W}^{(x)}$ とそれに対応するアクティベーション $\boldsymbol{H}^{(x)}$ との積に分解する．次に，楽曲 $\boldsymbol{Y}$ におけるクエリのアクティベーションを観測するために，固定したスペクトル $\boldsymbol{W}^{(x)}$ と非固定のスペクトル $\boldsymbol{W}^{(f)}$ とで構成される基底スペクトルを用いた NMF を適用する．非固定のスペクトル $\boldsymbol{W}^{(f)}$ は，クエリのフレーズに関与しない楽器音を表現するために必要な基底である．ここで，$\boldsymbol{W}^{(x)}$ と $\boldsymbol{W}^{(f)}$ に対応するアクティベーションをそれぞれ $\boldsymbol{H}^{(y)}$，$\boldsymbol{H}^{(f)}$ とする．最後に，クエリ単体のアクティベーション $\boldsymbol{H}^{(x)}$ と，混合音中のクエリのアクティベーション $\boldsymbol{H}^{(y)}$ との類似度を計算する．最後に，その類似度が大きな値を取る時刻をフレーズ登場時刻として検出する．

“ノンパラメトリック”な “ベイズ”NMF を用いなければならない重要な理由は 2 つある．

(1) 基底スペクトルの最適な数は，$\boldsymbol{X}$ および $\boldsymbol{Y}$ の複雑さに応じて自動的に決定されるべきであるから．基底の数を自動で決定するには，まず無限個の基底スペクトルが存在することを仮定し，観測データに合わせて有限個の基底のみが実質的にアクティベートされる機構を用いればよい．これは，ノンパラメトリックベイズ拡張された NMF を用いることで達成される．

(2) $\boldsymbol{H}^{(y)}$ と $\boldsymbol{H}^{(f)}$ とで異なる事前分布を与え，固定された基底 $\boldsymbol{W}^{(x)}$ が，制約のない基底スペクトル $\boldsymbol{W}^{(f)}$ よりも強調されるようにする必要があるから．事前分布を調節しないと，楽曲のスペクトログラム $\boldsymbol{Y}$ は，制約のない基底 $\boldsymbol{W}^{(f)}$ のみを用いて表現されてしまう恐れがある．本手法の特徴的な考えは，楽曲 $\boldsymbol{Y}$ を分解する際に，注目するフレーズが含まれていると「思い込んで」分解するということである．つまり，$\boldsymbol{Y}$ を表現するにあたって，できるだけ $\boldsymbol{W}^{(x)}$ という基底を使うように学習を誘導するというねらいがある．このような学習のねらいを反映するには，ベイジアンの枠組みを用いるのが自然である．

2.2 クエリ分解時の NMF

本手法では，ガンマ過程 NMF (GaP-NMF) [11] を利用し，$\boldsymbol{X}$ を非負値ベクトル $\boldsymbol{\theta} \in \mathbb{R}^{K_x}$ および 2 つの非負値行列 $\boldsymbol{W}^{(x)} \in \mathbb{R}^{M \times K_x}$ と $\boldsymbol{H}^{(x)} \in \mathbb{R}^{K_x \times N_x}$ の積によって近似する．より詳細には，クエリのスペクトログラム $\boldsymbol{X}$ は以下のように分解される:

$$X_{mn} \approx \sum_{k=1}^{K_x} \theta_k W^{(x)}_{mk} H^{(x)}_{kn}. \tag{1}$$

ここに θ_k は k 番目の基底の全体的な重み，$W^{(x)}_{mk}$ は k 番目の基底の周波数 m におけるパワー，$H^{(x)}_{kn}$ は k 番目の基底の時刻 n におけるアクティベーションである．$\boldsymbol{W}^{(x)}$ の各列は基底スペクトルを，$\boldsymbol{H}^{(x)}$ の各行は，その基底の時間

変化のパターンとなっている．

2.3 楽曲を分解するための半教師あり NMF

次に，楽曲のスペクトログラム $\boldsymbol{Y}$ を分解するための，半教師ありの NMF について説明する．本手法では，基底スペクトルの一部を $\boldsymbol{W}^{(x)}$ で置き換え固定する．学習のパラメータ更新時に基底 $\boldsymbol{W}$ を固定するという考え自体は既に Kirchhoff らが提案している [12]．しかし本手法は，$\boldsymbol{W}$ 全体のうち初めの K_x 列までの部分のみを固定する，という点で異なっている．

本手法では，楽曲 $\boldsymbol{Y}$ を近似するにあたって，固定した基底 $\boldsymbol{W}^{(x)}$ が広く使われるようにベイジアン NMF を定式化する．これを実現するために，$\boldsymbol{H}^{(y)}$ と $\boldsymbol{H}^{(f)}$ にそれぞれ異なるガンマ事前分布を与える．$\boldsymbol{H}^{(y)}$ のガンマ事前分布の形状母数は，$\boldsymbol{H}^{(f)}$ のものに比べてはるかに大きい値を取るように設定する．これは，ガンマ分布の期待値は，その形状母数に比例するためである．

2.4 アクティベーションの相関計算

上記の半教師あり NMF が完了した後に，クエリから得られたアクティベーション $\boldsymbol{H}^{(x)}$ と，楽曲から得たアクティベーション $\boldsymbol{H}^{(y)}$ との類似度を計算することで，フレーズの登場時刻を探す．もしクエリとほとんど同一のフレーズが楽曲中で奏でられたとすると，たとえ楽器の種類が異なっていても，クエリに類似するアクティベーションのパターンが表れると期待できる．具体的には，時刻 n における $\boldsymbol{H}^{(x)}$ と $\boldsymbol{H}^{(y)}$ との相関係数の和 $r(n)$ を以下の式によって求める:

$$r(n) = \frac{1}{K_x N_x} \sum_{k=1}^{K_x} \frac{\left(\boldsymbol{h}_{k1}^{(x)} - \overline{\boldsymbol{h}}_{k1}^{(x)}\right)^T \left(\boldsymbol{h}_{kn}^{(y)} - \overline{\boldsymbol{h}}_{kn}^{(y)}\right)}{\|\boldsymbol{h}_{k1}^{(x)} - \overline{\boldsymbol{h}}_{k1}^{(x)}\| \|\boldsymbol{h}_{kn}^{(y)} - \overline{\boldsymbol{h}}_{kn}^{(y)}\|}, \quad (2)$$

ここで,

$$\boldsymbol{h}_{ki}^{(\cdot)} = \left[H_{ki}^{(\cdot)} \cdots H_{k(i+N_x-1)}^{(\cdot)}\right]^T, \quad (3)$$

$$\overline{\boldsymbol{h}}_{kn}^{(\cdot)} = \frac{1}{N_x} \sum_{j=1}^{N_x} H_{k(n+j-1)}^{(\cdot)} \times [1 \cdots 1]^T. \quad (4)$$

最後に，時間軸上で相関係数のピークを求めることで，注目フレーズの開始点を検出する．ピーク検出は，以下の閾値処理に基づいて行った:

$$r(n) > \mu + 4\sigma. \quad (5)$$

ここで μ と σ は，全ての楽曲から得られた $r(n)$ の全体の平均および標準偏差を表す．

2.5 GaP-NMF の変分推論

本節では，あるスペクトログラム $\boldsymbol{V} \in \mathbb{R}^{M \times N}$ を，ノンパラメトリックベイズ NMF によって分解する際の推論方法について簡潔に述べる．詳細は Hoffman らの論文 [11] を参照されたい．まず，$\boldsymbol{\theta} \in \mathbb{R}^K$, $\boldsymbol{W} \in \mathbb{R}^{M \times K}$, $\boldsymbol{H} \in \mathbb{R}^{K \times N}$ は，それぞれある生成過程に従って確率的に生み出された値だと仮定する．ここでは，以下のようにガンマ分布を事前分布とした:

$$\begin{aligned} p(W_{mk}) &= \mathrm{Gamma}\left(a^{(W)}, b^{(W)}\right), \\ p(H_{kn}) &= \mathrm{Gamma}\left(a^{(H)}, b^{(H)}\right), \\ p(\theta_k) &= \mathrm{Gamma}\left(\frac{\alpha}{K}, \alpha c\right). \end{aligned} \quad (6)$$

ここに α は集中母数，K は混合音の構成要素数に比べ十分に大きい整数（理想的には無限の大きさ），c は $\boldsymbol{V}$ の平均値の逆数，つまり $c = \left(\frac{1}{MN}\sum_m \sum_n V_{mn}\right)^{-1}$ である．

次に，事後分布として以下の一般化逆ガウス分布（GIG 分布）を用いる:

$$\begin{aligned} q(W_{mk}) &= \mathrm{GIG}\left(\gamma_{mk}^{(W)}, \rho_{mk}^{(W)}, \tau_{mk}^{(W)}\right), \\ q(H_{kn}) &= \mathrm{GIG}\left(\gamma_{kn}^{(H)}, \rho_{kn}^{(H)}, \tau_{kn}^{(H)}\right), \\ q(\theta_k) &= \mathrm{GIG}\left(\gamma_k^{(\theta)}, \rho_k^{(\theta)}, \tau_k^{(\theta)}\right). \end{aligned} \quad (7)$$

これらのパラメータを推定するために，初めに ϕ_{kmn}, ω_{mn} という 2 つのパラメータについて，以下の式を用いて更新を行う．

$$\phi_{kmn} = \mathbb{E}_q\left[\frac{1}{\theta_k W_{mk} H_{kn}}\right]^{-1}, \quad (8)$$

$$\omega_{mn} = \sum_k \mathbb{E}_q\left[\theta_k W_{mk} H_{kn}\right]. \quad (9)$$

ϕ_{kmn}, ω_{mn} が計算できたら，次に GIG 分布のパラメータを以下の式によって更新する．

$$\gamma_{mk}^{(W)} = a^{(W)}, \quad \rho_{mk}^{(W)} = b^{(W)} + \mathbb{E}_q[\theta_k] \sum_n \frac{\mathbb{E}_q[H_{kn}]}{\omega_{mn}}, \quad \tau_{mk}^{(W)} = \mathbb{E}_q\left[\frac{1}{\theta_k}\right] \sum_n V_{mn} \phi_{kmn}^2 \mathbb{E}_q\left[\frac{1}{H_{kn}}\right], \quad (10)$$

$$\gamma_{kn}^{(H)} = a^{(H)}, \quad \rho_{kn}^{(H)} = b^{(H)} + \mathbb{E}_q[\theta_k] \sum_m \frac{\mathbb{E}_q[W_{mk}]}{\omega_{mn}}, \quad \tau_{kn}^{(H)} = \mathbb{E}_q\left[\frac{1}{\theta_k}\right] \sum_m V_{mn} \phi_{kmn}^2 \mathbb{E}_q\left[\frac{1}{W_{mk}}\right], \quad (11)$$

$$\gamma_k^{(\theta)} = \frac{\alpha}{K}, \quad \rho_k^{(\theta)} = \alpha c + \sum_m \sum_n \frac{\mathbb{E}_q[W_{mk} H_{kn}]}{\omega_{mn}}, \quad \tau_k^{(\theta)} = \sum_m \sum_n V_{mn} \phi_{kmn}^2 \mathbb{E}_q\left[\frac{1}{W_{mk} H_{kn}}\right]. \quad (12)$$

式 (8), (9) を計算するには，$\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{\theta}$ の期待値が必要となる．そこで，初めにそれらの期待値を正の乱数で初期化し，上記の式を用いて更新を繰り返していく．繰り返し回

数が増えるにつれて，ある正の整数 K^+ よりも大きい番号である k 番目の期待値 $\mathbb{E}_q[\theta_k]$ は 0 に近づいていく．そこで，$\mathbb{E}_q[\theta_k]$ の値が $\sum_k \mathbb{E}_q[\theta_k]$ に対して 60 dB を下回った場合，k 番目に対応する要素を削除することで，計算を速くすることができる．最終的に，実効的な基底の数 K^+ は，繰り返し計算の過程で徐々に減少していき，適切な基底数として自動的に決定されることになる．

3. マッチングを利用した従来手法

本章では，提案手法との比較のために，従来のマッチング手法を 2 種類説明する．1 つは音響特徴量のユークリッド距離を計算するもの（3.1 節），もう 1 つはスペクトログラム間の板倉–斎藤距離（IS ダイバージェンス）を計算する手法（3.2 節）である．

3.1 MFCC のユークリッド距離に基づくマッチング手法

フレーズが登場する時刻を，クエリと楽曲の短いセグメントとの音響特徴的な距離に注目して検出する手法である．本稿では，音声認識等で広く使用される，メル周波数ケプストラム係数（MFCC）を音響特徴量として用いた．具体的には，Auditory Toolbox Version 2 [13] を用いて，各フレームに対して 12 次元の特徴ベクトルを計算する．2 つの特徴ベクトル系列について，フレーム単位でユークリッド距離を算出し，クエリの長さ分だけ積算することにより最終的な距離を求める．

上記の距離を，クエリを 1 フレームずつ時間方向にずらして計算することを繰り返す．最後に，得られた距離の値が $m - s$ 未満の場合にフレーズを検出したものとみなす，という単純なピーク検出手法を用いた．ここで m, s はそれぞれ全フレームについての距離の平均，分散を表す．

3.2 板倉–斎藤距離に基づく DP マッチング手法

本節では，クエリ $\boldsymbol{X}$ と楽曲 $\boldsymbol{Y}$ との IS ダイバージェンスを直接計算することにより，フレーズ登場時刻を検出する手法について述べる．IS ダイバージェンスは，クエリのスペクトログラムが楽曲中に含まれている場合に，通常のユークリッド距離やカルバック–ライブラ（KL）ダイバージェンスといった距離尺度に比べ，ペナルティがより小さくなる尺度であるため，IS ダイバージェンスを用いることは理論的に妥当である．

フレーズ開始時刻を効率的に探すために，本手法では IS ダイバージェンスに基づく動的計画法（DP）マッチングを導入する．初めに，距離行列 $\boldsymbol{D} \in \mathbb{R}^{N_x \times N_y}$ を計算する．$\boldsymbol{D}$ の各要素 $D(i,j)$ は，$\boldsymbol{X}$ の i 番目のフレームと，$\boldsymbol{Y}$ の j 番目のフレームとの IS ダイバージェンスの値になる（$1 \le i \le N_x$ および $1 \le j \le N_y$）．$D(i,j)$ は以下の式によって計算される:

$$D(i,j) = \mathcal{D}_{\mathrm{IS}}(\boldsymbol{X}_i|\boldsymbol{Y}_j) = \sum_m \left(-\log \frac{X_{mi}}{Y_{mj}} + \frac{X_{mi}}{Y_{mj}} - 1 \right). \tag{13}$$

ただし，m は周波数方向の要素番号である．次に，累積距離行列 $\boldsymbol{E} \in \mathbb{R}^{N_x \times N_y}$ を求める．まず $\boldsymbol{E}$ を，全ての j について $E(1,j) = 0$，全ての i について $E(i,1) = \infty$ と初期化する．累積距離行列の各要素 $E(i,j)$ は，以下の式を用いて次々に求められる:

$$E(i,j) = \min \left\{ \begin{array}{l} 1)\ E(i-1,j-2) + 2D(i,j-1) \\ 2)\ E(i-1,j-1) + D(i,j) \\ 3)\ E(i-2,j-1) + 2D(i-1,j) \end{array} \right\} + D(i,j). \tag{14}$$

結果的に，クエリと，楽曲中の j 番目のフレームを終端とするセグメントとの距離 $E(N_x, j)$ が求められる．また，累積コスト行列を $\boldsymbol{C} \in \mathbb{R}^{N_x \times N_y}$ とし，上式の 1) – 3) の 3 つの場合について，それぞれ以下のように計算する:

$$C(i,j) = \left\{ \begin{array}{l} 1)\ C(i-1,j-2) + 3 \\ 2)\ C(i-1,j-1) + 2 \\ 3)\ C(i-2,j-1) + 3 \end{array} \right. \tag{15}$$

ただし，$\boldsymbol{C}$ の各要素は 0 で初期化する．

これらの式は，楽曲中のフレーズの長さが，クエリの長さの半分から 2 倍までの範囲にあることを許容するものである．

フレーズ検出点は，正規化された累積距離 $\frac{E(N_x,j)}{C(N_x,j)}$ の極小値として求められる．具体的には，得られた距離の値が $M - S/5$ となる時刻を，検出されたフレーズの終了時刻であるとみなす．ただし，M および S はそれぞれ，全曲から得られた正規化累積距離の中央値，標準偏差を表す．距離に閾値処理を施す際に平均でなく中央値を使う理由は，累積距離はまれに極端に大きい値（異常値）をとるためである．距離の平均は，このような異常値の影響を受け，大きすぎる値を取る傾向にある．また，S を計算するにあたって，実用上の理由から，10^6 を超える値を無視することにする（$\frac{E(N_x,j)}{C(N_x,j)}$ のほぼすべての値は 10^3 から 10^4 の範囲に収まる）．フレーズの終端が検出されると，その点から経路をバックトレースすることにより，フレーズの開始点も簡単に求めることができる．

4. 実験

本章では，2 章の提案手法，および 3.1 節，3.2 節の従来手法について，Query-by-Phrase (QBP) の精度を評価する比較実験について述べる．

4.1 実験条件

提案手法および 2 つの従来手法それぞれについて，3 つの異なる実験条件の下で評価を行った:

表 1 楽曲中と全く同じフレーズをクエリとした場合の実験結果

	Precision (%)	Recall (%)	F-measure (%)
MFCC	24.8	35.0	29.0
DP	0	0	0
提案手法	43.2	88.0	57.9

表 2 楽曲中とは異なる楽器でクエリを演奏した場合の実験結果

	Precision (%)	Recall (%)	F-measure (%)
MFCC	0	0	0
DP	2.1	21.7	3.9
提案手法	26.9	56.7	36.5

表 3 楽曲よりも 20% 速くクエリを演奏した場合の実験結果

	Precision (%)	Recall (%)	F-measure (%)
MFCC	0	0	0
DP	0	0	0
提案手法	15.8	45.0	23.4

(1) クエリと全く同一のフレーズが楽曲に含まれている場合（exact-match）.
(2) クエリが楽曲中とは異なる種類の楽器である場合（音色変化）.
(3) クエリが楽曲よりも速いテンポで演奏された場合（テンポ変化）.

本実験では，RWC 研究用音楽データベース (ポピュラー音楽) [14] から 4 曲 (No.1, No.19, No.42, No.77) を選びデータベース楽曲群とした．クエリについては，以下の 50 種類を準備した:

(1) 元の 4 つの楽曲のマルチトラック収録の一部を切り抜いて作ったもの (10 個).
(2) 元の楽曲とは異なる楽器で演奏したクエリ (30 個，著者らの手動の演奏により作成).
(3) 元の楽曲と同じ楽器だが，演奏のテンポを 20% 速めたもの (10 個).

各クエリは単一の楽器の演奏であり，演奏時間は 4〜9 秒である．これらのフレーズは，必ずしも元の楽曲中で目立っているわけではない（主旋律ばかりではない）ことに注意されたい．各音響信号は 16 kHz でサンプリングされたモノラル信号とし，各信号に対し 10 ms 間隔で短時間フレームをシフトさせながらウェーブレット変換を施す．短時間フーリエ変換（STFT）を用いずウェーブレット変換を採用した理由は，低周波数帯における時間分解能を高くするためである．ガボールウェーブレット関数の標準偏差は 3.75 ms (サンプリング点 60 点に相当) とした．周波数方向の間隔は 10 cent であり，周波数の値の範囲は 900〜10720 cent である．

クエリを NMF を用いて分解する際の超パラメータは $\alpha = 1$, $K = 100$, $a^{(W^{(x)})} = b^{(W^{(x)})} = a^{(H^{(x)})} = 0.1$, $b^{(H^{(x)})} = c$ とした．一方，半教師あり学習を用いて NMF を計算する際には，$a^{(W^{(x)})} = b^{(W^{(x)})} = a^{(W^{(f)})} = b^{(W^{(f)})} = 0.1$, $a^{(H^{(y)})} = 10$, $a^{(H^{(f)})} = 0.01$, $b^{(H^{(y)})} = b^{(H^{(f)})} = c$ とした．尺度パラメータの逆数 $b^{(H)}$ の値は，対象の音響信号の経験的尺度によって調節される．また，$a^{(\cdot)}$ の値を小さくすることで，無限空間においてよりスパースな学習へと誘導できることにも注意されたい．

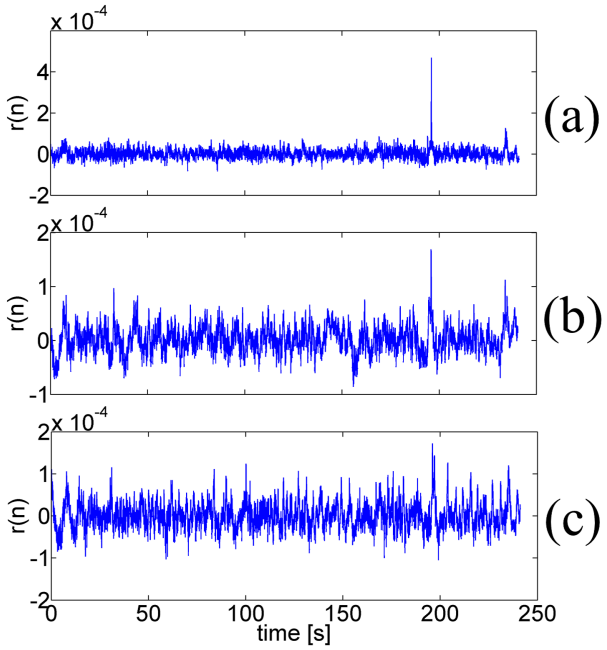

図 2 相関係数の和 $r(n)$ のグラフ．検索対象楽曲は RWC-MDB-P-2001 No.42. (a) 検索対象となるサックスのフレーズと全く同じ音をクエリにした場合. (b) クエリの音色をストリングスにした場合. (c) クエリの演奏テンポを 20%速くした場合.

各手法の QBP 精度を測るために，情報検索分野で広く用いられる F 尺度の平均を求める．適合率 (precision rate) は，検索結果として得たすべてのフレーズのうち，正しく得られたものの割合として定義される．一方，再現率 (recall rate) は，検索対象となるデータベース中に存在する正解フレーズのうち，検索結果として得られたものの割合である．なお本実験では，各クエリのフレーズはある 1 曲のみに登場する（1 曲に複数回登場することはあり得る）．再現率 P と適合率 R を求めた後に，F 尺度 F を $F = \frac{2PR}{P+R}$ という式によって計算する．

4.2 実験結果

3 つの手法の検索精度の平均は表 1–3 の通りとなった．提案手法は，従来の手法に比べて QBP 精度がはるかに優れていることを確認できた．図 2 は，実際にあるクエリのフレーズ（楽曲中ではサックスの音色）が含まれる楽曲に対し，計算された相関係数の和 $r(n)$ をグラフで示したものである．これを見ても，本手法によりクエリのフレーズの開始時刻が正しく検出されることが分かる．MFCC に基づく手法でも，exact-match の条件においては正しく検索できる場合もあるが，音色変化やテンポ変化に対して頑健ではない．一方，DP マッチング手法については，IS ダイバージェンスの値は音響特徴的な類似度よりもむしろ音

量変化に非常に敏感であるため，ほとんどの場合について正しい検出ができなかった．正しい位置にコスト関数の極小値が現れる場合もあるが，混合音のスペクトログラムから明確なフレーズ終了位置を検出することが難しいため，それらの極小値も顕著なものではなくなってしまう．

提案手法は 3 つの中で最も精度が高かったものの，実用化に向けては精度の改善が求められる．主な課題は，適合率が再現率に比べて低くなることである．誤った位置が検出されるのは，クエリがスタッカート [*1] などの奏法で演奏されている場合，スタッカート音の発音間隔分だけ時間方向に前後した箇所に誤ったピークが生じることなどが原因である．

5. おわりに

本稿では，クエリとして与えられたフレーズが楽曲中に登場する時刻を検出するための Query-by-Phrase (QBP) 手法を提案した．本手法は，音楽音響信号を完全に採譜することなしに，半教師あり非負値行列因子分解（NMF）を用いることでクエリと楽曲の構成要素の一部との距離を適切に計算することができる．実験結果から，本手法が従来のマッチング手法よりも優れた検索精度を示すことが分かった．さらに，クエリとなるフレーズが異なる楽器で演奏された場合（音色変化）や，より速いテンポで演奏された場合（テンポ変化）にも，正しくフレーズ登場位置を検出できる可能性があることも確認できた．

今後は，上記のような音色変化やテンポ変化が生じた際の精度を改善するため，クエリの基底スペクトルを，ギター固有のノイズ成分など楽器に依存する基底と，調波構造をもつ楽器に共通の基底とに分類すること，普遍的な基底スペクトルの組を予め用意することなどを検討している．また，ノンパラメトリックベイズ NMF に基づく本手法の計算コストの削減も重要な課題である．

謝辞 本研究は JST CREST「OngaCREST プロジェクト」の支援を受けた．

参考文献

[1] Li T. and Ogihara M.: Content-based Music Similarity Search and Emotion Detection, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5:705–708, 2004.

[2] Logan B. and Salamon A.: A Music Similarity Function Based on Signal Analysis, *International Conference on Multimedia and Expo (ICME)*, pp. 745–748, 2001.

[3] Ramona M. and Peeters G: AudioPrint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 818–822, 2013.

[4] Haitsma J. and Kalker T.: A Highly Robust Audio Fingerprinting System, *International Conference on Music Information Retrieval (ISMIR)*, pp. 107–115, 2002.

[5] McNab R. J., Smith L. A., Witten I. H., Henderson C. L., and Cunningham S. J.: Towards the Digital Music Library: Tune Retrieval from Acoustic Input, in *Proceedings of the first ACM international conference on Digital libraries*, pp. 11–18, 1996.

[6] Ghias A., Logan J., Chamberlin D., and Smith B. C.: Query By Humming: Musical Information Retrieval in an Audio Database, *Proceedings of the third ACM international conference on Multimedia*, pp. 231–236, 1995.

[7] Nishimura T., Hashiguchi H., Takita J., Zhang J. X., Goto M. and Oka R.: Music Signal Spotting Retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming, *International Conference on Music Information Retrieval (ISMIR)*, pp. 211–218, 2001.

[8] Kameoka H., Ochiai K., Nakano M., Tsuchiya M. and Sagayama S.: Context-free 2D Tree Structure Model of Musical Notes for Bayesian Modeling of Polyphonic Spectrograms, *International Conference on Music Information Retrieval (ISMIR)*, pp. 307–312, 2012.

[9] Grindlay G. and Ellis D. P. W.: A Probabilistic Subspace Model for Multi-instrument Polyphonic Transcription, *International Conference on Music Information Retrieval (ISMIR)*, pp. 21–26, 2010.

[10] Ryynänen M. and Klapuri A.: Automatic Bass Line Transcription from Streaming Polyphonic Audio, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4:1437–1440, 2007.

[11] Hoffman M. D., Blei D. M. and Cook P. R.: Bayesian Nonparametric Matrix Factorization for Recorded Music, *International Conference on Machine Learning (ICML)*, pp. 439–446, 2010.

[12] Kirchhoff H., Dixon S. and Klapuri A.: Shift-variant Non-negative Matrix Deconvolution for Music Transcription, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 125–128, 2012.

[13] Slaney M.: Auditory Toolbox Version 2, *Technical Report #1998-010*, Interval Research Corporation, 1998.

[14] Goto M., Hashiguchi H., Nishimura T. and Oka R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *International Conference on Music Information Retrieval (ISMIR)*, pp. 287–288, 2002.

[*1] スタッカート：1 つ 1 つの音を短く切って演奏すること．

付　　録

A.1 確率分布の式

ガンマ分布

$$\mathrm{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \qquad \text{(A.1)}$$

$\Gamma(x)$: ガンマ関数

一般化逆ガウス分布

$$\mathrm{GIG}(x|a,b,c) = \frac{(b/c)^{\frac{a}{2}} x^{a-1}}{2\mathcal{K}_a(2\sqrt{bc})} e^{-(bx+\frac{c}{x})} \qquad \text{(A.2)}$$

$\mathcal{K}_\nu(x)$: 第 2 種変形ベッセル関数