Vol.2018-MUS-118 No.17
Vol.2018-SLP-120 No.17
2018/2/21

カバーソング同定法を応用したメドレー楽曲における楽曲断片検出法の提案

佐藤 僚太[1,a)]　竹川 佳成[2,b)]　平田 圭二[2,c)]

概要：本稿ではカバーソング同定法（Cover Song Identification, CSI）によるメドレー楽曲における楽曲断片検出手法について述べる．本研究で扱うメドレー楽曲は，音楽的展開を考慮し楽曲断片を編曲して連結するため，ひとつの楽曲であるかのように聴こえる．そのため，楽曲の変わり目を認識するためには，メドレー楽曲から楽曲断片を検出することが必要である．我々は，編曲された楽曲を同定する CSI 手法を用いてメドレー楽曲中から楽曲断片を検出する手法を提案する．本手法では音楽的構造を考慮するため，2 曲の類似度行列である CRP(Cross Recurrence Plots) 行列を作成し，CSI を行う Serrà らの手法 [10] を用いた．CRP 行列からカバーソングを定量的に評価するための累積値行列 Q を作成し，$Q_{max} = max\{Q_{i,j}\}$ を楽曲間比較することで楽曲断片の検出を試みる．実験結果から，一部のメドレー楽曲において本手法の有用性が示された他，メロディ抽出の精度の向上によって本手法の楽曲断片検出の精度も向上することが示唆された．

Detection Method of Musical Segments based on Cross Recurrence Quantification for Cover Song Identification

Ryota Sato[1,a)]　Yoshinari Takegawa[2,b)]　Keiji Hirata[2,c)]

1. はじめに

聴き手が楽曲をブラウジングすること（能動的音楽鑑賞 [3]）を目的としたメドレー楽曲の聴取がされている．Customer Generated Media (CGM) サービスの niconico において 100 万回以上再生された「音楽」カテゴリ 246 本の動画の内，13 本を「ニコニコメドレーシリーズ」のタグが付与されたメドレー楽曲の動画が占めている．メドレー楽曲とは，編曲された複数の楽曲断片を連結して作られる新たな形式の楽曲を指す．ある楽曲が終わって次の楽曲が再生されるプレイリストのような形式に対し，メドレー楽曲では音楽的展開が考慮されているため，ひとつの楽曲の聴取であるかのように複数楽曲の聴取が行われる．

しかし，メドレー楽曲における楽曲の変わり目（楽曲遷移点）が，楽曲が変わったことに気づかないほど自然なものであるため，聴き手が楽曲遷移点を検出するのが困難であるという問題がある．特に，楽曲遷移点の前後が聴き手の未知楽曲である時の検出が困難であり，第三者の存在無しにはメドレー楽曲を構成している全ての楽曲を網羅することが難しい．そのため現状では，メドレー楽曲の作曲者もしくは聴き手が楽曲遷移点の情報を wiki などで提供している．メドレー楽曲の音響情報から自動で楽曲遷移点を検出することが実現すれば，メドレー楽曲のどの部分にどの楽曲が使われているかを第三者に頼ること無く知ることができ，聴き手の未知楽曲に対する能動的音楽鑑賞をより促すことができる．

そこで本論文では，メドレー楽曲から楽曲遷移点を検出することを目指す．聴き手が楽曲遷移点を検出するプロセスにおいて，その前後が何の楽曲のどの部分かを同定する，楽曲断片の検出が必要である．また，メドレー楽曲がひと

[1] 公立はこだて未来大学大学院
Graduate School of Future University Hakodate
[2] 公立はこだて未来大学
Future University Hakodate
[a)] g2117024@fun.ac.jp
[b)] yoshi@fun.ac.jp
[c)] hirata@fun.ac.jp

つの楽曲であるかのような音楽的展開をするために，楽曲断片同士を連結する際，楽曲に対してテンポ，ピッチ，構成楽器，和音の変更，旋律音の加減算等の編曲がされている．そのため，Wang の音声指紋を用いたマッチング手法 [11] のような，原曲同士のマッチングを対象とした手法を用いることが難しい．また，Foote の Self-Similarity Matrix[2] や Dannenberg ら [1] の楽曲構造分析では，メドレー楽曲中に繰り返し構造が出現することが少なく，多くの場合で楽曲遷移点を検出することが困難である．そのため，楽曲断片の検出を実現するために，編曲された楽曲と原曲とのマッチングを考慮した手法を用いることが望ましい．

本論文ではこの問題を解決するため，Serrà らの相互再帰定量化 (Cross Recurrence Quantification, CRP) によるカバーソング同定法 (Cover Song Identification, CSI)[10] を用いて楽曲断片検出手法の提案を行う．CSI のために提案される手法は原曲とそのカバーソングとのマッチングを目的としているため，編曲による影響を考慮した楽曲断片の検出を行うことができる．また CRP を用いることで，音楽的構造のまとまりが考慮できる他，2 曲のマッチングにおける全探索の計算量を削減することができる．本論文では CRP を用いた手法に加え，メドレー楽曲中のある時刻において，構成する楽曲は 1 曲のみであるという特徴を用いて，楽曲断片を検出する精度の向上を図った．

2. 関連研究

2.1 SoundCompass

楽曲の区間の一部分から楽曲認識を行う関連研究に，小杉らの SoundCompass[4] が挙げられる．このシステムは，ユーザがハミングした楽曲の一部分から楽曲名を提示するものである．ユーザが曖昧な場所から歌い出すことに対応するため，メロディの MIDI データを開始地点をずらしながら複数に分割し，楽曲検索のためのデータベースを構築している．入力されたハミングとデータベースからのような音高の特徴ベクトルをそれぞれ生成し，ベクトル同士の類似度を楽曲の類似度とすることで，曖昧な入力による楽曲認識を実現した．メロディの MIDI データを分割することによるデータベースの構築は，構築されるデータベースが非常に多くの冗長なデータを含むという問題点が指摘されている．また，データベースとなるメロディの抽出とその整形は人手によって行われるため，データベースを 1 曲増やすためにも多くの時間がかかる．

2.2 Query by Phrase

ガンマ過程非負値行列因子分解（Gamma Process Non-negative Matrix Factorization, GaP-NMF）を用いて，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による音楽音響信号からのフレーズ検出手法が増田らによって提案されている [6]．GaP-NMF によって得られた楽曲とクエリのアクティベーションの相関からフレーズの開始時刻を検出している．この手法の重要な課題は計算コストの削減であることが指摘されている．GaP-NMF の計算コストに加え，フレーズの開始時刻検出による相関計算を楽曲全体に対して行う必要があるなど，膨大な時間の計算コストがかかることが予想される．

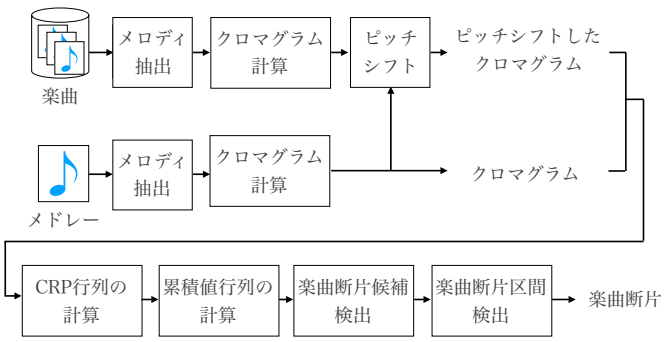


図 1 楽曲断片検出システム構成図

3. メドレー楽曲における楽曲断片検出手法

本章ではメドレー楽曲から楽曲断片を検出する提案手法ついて説明する．提案手法の構成を図 1 に示す．本手法では，Serrà らの CSI 手法に基づき，メドレー楽曲とその構成楽曲のクロマグラムに対して Optimal Transposition Index(OTI) の計算を行い，CRP 行列と累積値行列 Q を作成し，行列 Q の最大値とそのインデックスを得る．このとき，メドレー楽曲とその構成楽曲から melodia[7] を用いてメロディを抽出し，クロマグラムを計算する．そして，行列 Q から得られた最大値とそのインデックスから，楽曲断片候補を生成する．この検出される楽曲断片候補は，不完全なメロディ抽出等の要因によって他の楽曲と区間が重複してしまうことがほとんどである．そのため，楽曲断片の区間を一意に決めるために，他の楽曲の結果と比較しながら楽曲断片を検出し出力する．また，楽曲断片検出において本手法では以下の前提条件を設定した．

- メロディ，ハーモニー，リズムの音楽の三要素のうち，メロディが最も編曲による影響を受けにくい．
- メドレー楽曲中では複数の楽曲が同じ時刻において存在しない．

3.1 Marwan らの CRP 行列作成手法

本節ではまず，Marwan らの提案した 2 つの異なる信号から CRP 行列を作成する手法 [5] について説明する．CRP 行列は，2 つ信号のの異なる時刻における類似度行列である．

$$CR_{i,j} = \Theta(\epsilon_i^x - ||x_i - y_j||)\Theta(\epsilon_j^y - ||x_i - y_j||) \quad (1)$$

ここで，for $i = 1, ...N_x$, for $j = 1, ..., N_y$ である．N_x は信号 X のベクトルの総数を表しており，N_y についても同様である．x_i は，信号 X の時刻 i におけるベクトルを表しており，y_j についても同様である．ϵ_i^x は，N_x 個の x_i ベクトルの内，類似度の似ている上位 κ%の閾値を表している．閾値を超えていれば非ゼロ数となり，超えていなければ 0 の値をとる．ϵ_j^y についても同様である．$\Theta(\cdot)$ は Heviside の階段関数であり，以下のような値をとる．

$$\Theta(v) = \begin{cases} 0 & if\ v < 0 \\ 1 & otherwise \end{cases} \quad (2)$$

CRP 行列は座標 (i, j) における，2 つの信号のある時刻 i とある時刻 j の類似度が閾値より高いかを表す値で構成されている．

3.2 CSI のための CRP 行列の作成

本節では，CRP 行列作成手法を CSI に用いる方法 [10] に基づいた，本手法における楽曲から CSI を行うための CRP 行列を作成する方法について説明する．

まず，各楽曲からメロディを抽出する．Serrà らの手法では，楽曲から直接クロマグラムを計算していたのに対し，本手法では，抽出したメロディからクロマグラムを計算する点において異なる．これは，メロディ，ハーモニー，リズムの音楽の三要素のうち，メロディが最も編曲による影響を受けにくいという前提条件から，本手法の CSI における楽曲同定にメロディの情報のみを扱うこととしたためである．抽出したメロディ情報を楽曲信号として CRP 行列作成に用いる．

次に楽曲信号からクロマグラムを抽出し，OTI の計算を行う．まず，楽曲信号 A, B からクロマグラム $\overrightarrow{h_A}$，$\overrightarrow{h_B}$ を得る．このクロマグラムから Serrà らの手法 [8][9] に基づき，楽曲信号全体の正規化したヒストグラム

$$\overrightarrow{g_A} = \frac{\sum_{i=1}^{N} \overrightarrow{h_{A,i}}}{max\{\sum_{i=1}^{N} \overrightarrow{h_{A,i}}\}} \quad (3)$$

を得る．ここで，N は楽曲信号 A の総フレーム数である．楽曲信号の B についても同様に計算し $\overrightarrow{g_B}$ を得る．そして，$\overrightarrow{g_A}$，$\overrightarrow{g_B}$ の OTI は以下の計算で求められる．

$$OTI(\overrightarrow{h_A}, \overrightarrow{h_B}) = \underset{0 \le id \le N_H - 1}{\arg\max} \{\overrightarrow{h_A} \cdot circshift_R(\overrightarrow{h_B}, id)\} \quad (4)$$

ここで，N_H はクロマグラムのピッチクラス数を表す．本手法では $N_H = 12$ とした．また，$circshift_R(\overrightarrow{h}, id)$ は，図 2 のように $\overrightarrow{h}$ のインデックスを id 個ずらす処理を表す．最後に，計算した OTI を用いて以下のように $\overrightarrow{h_A}$ のインデックスをずらす．

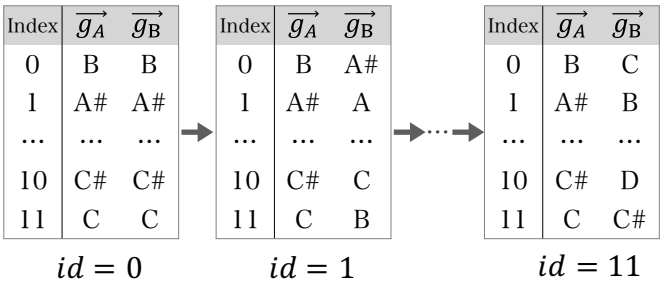


図 2 $circshift$ 関数

$$\overrightarrow{h_{A,i}^{Tr}} = circshift_R(\overrightarrow{h_{A,i}}, OTI) \quad (5)$$

以上のように楽曲信号から得たクロマグラムのインデックスを OTI によって揃えた $\overrightarrow{h_A^{Tr}}$ ものを入力として，3.1 節の手法で CRP 行列 CR を作成する．

3.3 累積値行列 Q の作成

3.1 節，3.2 節の手法によって作成した行列 CR から，CSI を行うため，Dynamic Time Warping(DTW) アルゴリズムに類似したアルゴリズムで計算した累積値行列 Q を計算する．

行列 Q は，$Q_{1,j} = Q_{2,j} = Q_{i,1} = Q_{i,2} = 0\ for\ 1, ..., N_x,\ j = 1, ..., N_y$ と初期化された後，以下のように作成する．

$$Q_{i,j} = \begin{cases} max\{Q_{i-1,j-1}, Q_{i-2,j-1}, Q_{i-1_j-2}\} + 1 & if\ CR_{i,j} = 1 \\ max\left\{ \begin{array}{c} 0, \\ Q_{i-1,j-1} - \gamma(CR_{i-1,j-1}), \\ Q_{i-2,j-1} - \gamma(CR_{i-2,j-1}), \\ Q_{i-1,j-2} - \gamma(CR_{i-1,j-2}) \end{array} \right\} & if\ CR_{i,j} = 0 \end{cases} \quad (6)$$

ここで，for $i = 3, ...N_x$, for $j = 3, ..., N_y$ である．$\gamma(\cdot)$ は，行列 CR の要素が類似していなかった場合に，メロディの加減算を考慮し，累積値の減少を抑えるためのパラメータであり，以下のような値をとる．

$$\gamma(z) = \begin{cases} \gamma_o & if\ z = 1 \\ \gamma_e & if\ z = 0 \end{cases} \quad (7)$$

$z = 1$ の場合，前の要素が類似しているため，音の加減算が行われた可能性がある．そのため，γ_o では累積値の減少幅を少なく設定する．$z = 0$ の場合，類似していない要

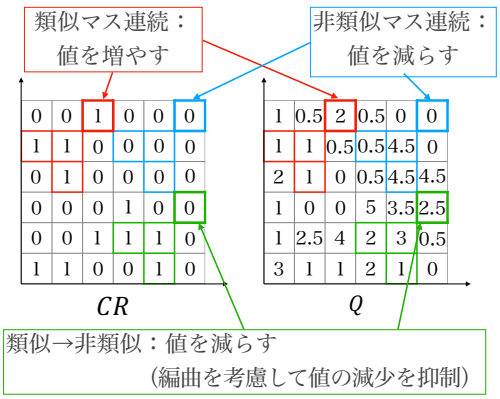


図 3 行列 Q

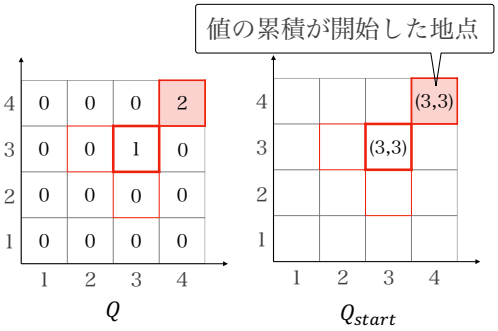


図 4 行列 Q_{start}

素が連続したため，この要素の付近は類似していない区間である可能性がある．そのため誤検出を防ぐために，γ_o では累積値の減少幅を多く設定する．γ_o，γ_e のパラメータは Serrà らによって検証されており，CSI においては $\gamma_o = 5.0$，$\gamma_e = 0.5$ を用いるのが適当であるとしている．

作成された行列 Q から，$Q_{max} = max\{Q_{i,j}\}$ となる Q_{max} の比較によって，CSI を行うことができる．合わせて，類似区間の終了地点が (i,j) であると検出することができる．

3.4 楽曲断片候補の検出

本節では，先行研究によって作成された行列 Q を用いて，メドレー楽曲から楽曲断片候補を検出するための手法について説明する．3.3 節において，Q_{max} となる地点 (i,j) が類似区間の終了地点であることを述べた．類似区間の開始地点を求めることが可能であれば，その開始地点から終了地点を楽曲断片として検出することができる．本手法では楽曲断片候補を作成するために，楽曲断片の開始地点を求めるための，開始地点候補行列 Q_{start} を作成する．CSI の比較のために用いる行列 Q の値は，最も類似している直前の値が累積するアルゴリズムである．つまり，どの地点の値を参照したかを遡ることで，楽曲中の類似している区間の開始地点を求めるため，Q の作成と並行して以下のように Q_{start} 行列を作成する（図 4）．

$(i) CR_{i,j} = 1$ のとき

$$Q_{start(i,j)} = \begin{cases} (i,j) & if\ max\{C\} = 0.0 \\ \arg\max_{i,j}\{C\} & otherwise \end{cases} \tag{8}$$

ここで，$C = \{Q_{i-1,j-1}, Q_{i-2,j-1}, Q_{i-1_j-2}\}$ である．累積値が 0.0 のとき，類似している (i,j) 地点が区間の開始地点となりうる．それ以外の場合は，(i,j) の開始地点を，累積する値をもつ地点の開始地点とする．

$(ii) CR_{i,j} = 0$ のとき

$$Q_{start(i,j)} = \begin{cases} \phi & if\ max\{C\} = 0.0 \\ \arg\max_{i,j}\{C\} & otherwise \end{cases} \tag{9}$$

ここで，$C = \{0, Q_{i-1,j-1}, Q_{i-2,j-1}, Q_{i-1_j-2}\}$ である．累積値が 0.0 のとき，(i,j) 地点も区間の開始地点となりえない．それ以外の場合は，それ以外の場合は，(i,j) の開始地点を，累積する値をもつ地点の開始地点とする．

行列 Q_{start} は以上のように定義され，$max\{Q_{i,j}\}$ となる (i,j) が類似区間の終了地点であるとき，その開始地点は $Q_{start(i,j)}$ である．

3.5 楽曲断片区間の決定

本節では，3.4 節で検出した楽曲断片候補を用いて，メドレー楽曲上における区間を決定するための手法を述べる．

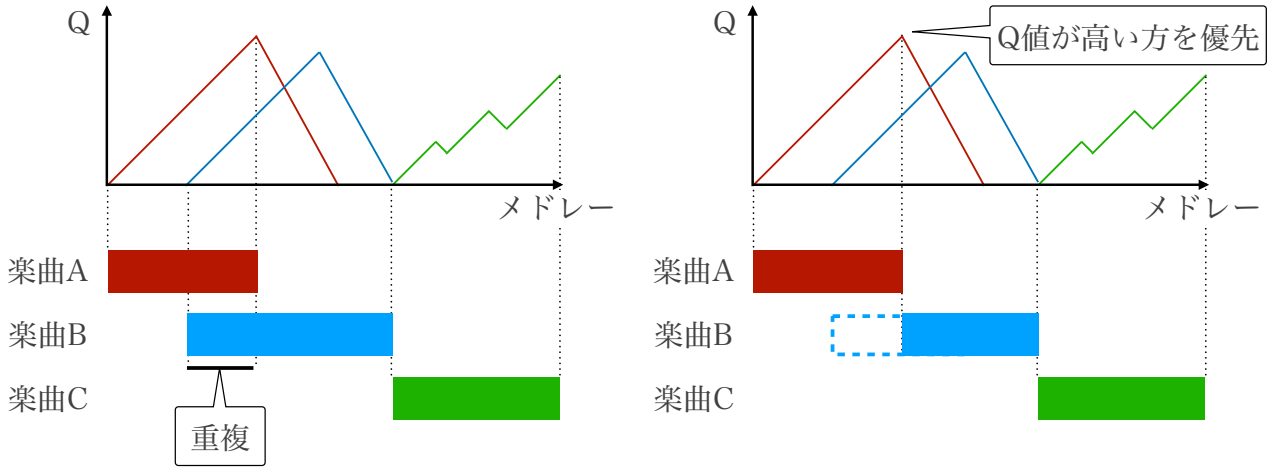
本手法では，Q_{max} の高い楽曲を優先的にメドレー楽曲上における楽曲断片であるとして，楽曲断片区間の決定を行う．楽曲断片の検出例を図 5(a) に示す．検出された楽曲断片候補は，不完全なメロディ抽出や類似した他楽曲とのマッチングによって，不正確な検出となり，他楽曲断片と区間が重複することがある．そこで，メドレー楽曲上の時刻において，楽曲が一意に決まるよう，最も高い Q_{max} の値を持つ楽曲を，その時刻における楽曲断片であるとした（図 5(b)）．

4. 評価実験

メドレー楽曲の楽曲断片の検出について，提案手法の精度評価を行った．

4.1 実験条件

構成楽曲数が 5〜10 曲かつ 2 分以下のメドレー楽曲と，その構成楽曲のホモフォニックなデータを用いた．楽曲断片の開始時刻・終了時刻の正解データは，実験で用いるホモフォニックなメドレー楽曲のデータから単旋律を MIDI トラックから抽出し，MIDI 上で楽曲ごとに分割した時刻



(a) 楽曲断片に重複がある例　　(b) 楽曲断片の重複をなくした例

図 5　楽曲断片の検出例

をそれぞれ正解データとした．

CRP 行列を作成する際のパラメータは，$k = 0.1$，行列 Q を作成する際のパラメータは $\gamma_o = 5.0$, $\gamma_e = 0.5$ とした．また，melodia[7] によるメロディ抽出のパラメータは，voicing= 0.2, minfqr= 55.0, maxfqr=1760.0 とした．

検出された楽曲断片を正解データと比較し，F 値を用いて精度評価を行う．

4.2　実験結果

実験による各メドレー楽曲に対する精度を表 1 に示す．ー は構成楽曲が少ないため，ID に対応する楽曲が存在しないことを表す．

F 値が最も高いメドレー楽曲は 86.48%であった（メドレー楽曲 ID1）．対して F 値が最も高いメドレー楽曲は 13.82%であった（メドレー楽曲は ID2）．また，正しい楽曲断片が 1 つも検出できず，0.00%となった検出結果もいくつか見られた．

特に精度の低かったメドレー楽曲 ID2 の詳細な精度の結果を表 2 に示す．結果から，楽曲 ID1, 2 の適合率が 100.00%であるにも関わらず，再現率が極端に低いため，F 値が非常に低くなっていることがわかる．また，楽曲 ID5 について，再現率が 97.87%であるにも関わらず，適合率が低いため，F 値の低い結果となった．

表 1　楽曲断片検出の F 値 (%)

		メドレー楽曲 ID					
		1	2	3	4	5	平均
楽曲 ID	1	99.18	12.56	98.98	0.00	87.44	
	2	82.27	15.33	0.00	0.00	50.26	
	3	58.44	0.00	61.30	67.43	64.03	
	4	71.63	0.00	45.52	0.00	97.03	
	5	96.08	41.22	50.34	82.43	87.18	
	6	98.17	ー	ー	67.10	ー	
	7	99.56	ー	ー	ー	ー	
平均		86.48	13.82	51.23	43.39	77.19	53.22

表 2　メドレー楽曲 ID2 の精度 (%)

		適合率 (%)	再現率 (%))	F 値 (%)
楽曲 ID	1	100.00	6.71	12.56
	2	100.00	8.33	15.53
	3	0.00	0.00	0.00
	4	0.00	0.00	0.00
	5	26.11	97.87	41.22

4.3　結果に対する考察

本手法による楽曲断片検出に対する結果から，精度が低くなってしまう原因として以下のような理由が考えられる．

楽曲断片検出手法が不適切

表 2 のように，適合率と再現率が極端に高く出てしまう結果から，提案した楽曲断片検出手法が不適切である可能性がある．メドレー楽曲 ID2 のシステム出力は，楽曲 ID5 の曲がであるという結果が多くを占めており，その他の楽曲が出力が極端に少なかったためこのような結果となった．これは，楽曲断片候補の検出した区間が正解データに比べて長すぎたもしくは，楽曲断片区間の決定において，適切に各楽曲の Q_{max} の値が比較できていない可能性があることが考えられる．

メロディ抽出の精度が不十分

また，メロディ抽出の精度が不十分である可能性も考えられる．人手によってメロディを MIDI 情報で作成し，音響信号として出力したものに本手法を適用して，同様に楽曲断片検出を行った結果を表 3 に示す．ほぼ全ての結果において有効な精度で楽曲断片が検出できていることが読み取れる．現在は melodia を用いてメロディ抽出を行っているが，適切なパラメータ調整

表 3　人手で作られたメロディによる楽曲断片検出の F 値 (%)

	メドレー楽曲 ID					
	1	2	3	4	5	平均
平均	98.94	90.53	92.48	85.86	97.66	93.09

や，アルゴリズムを別のものに変えることで，本手法による楽曲断片の検出精度が向上する可能性があるといえる．

5. おわりに

本論文では CRP 行列を用いた CSI 手法によるメドレー楽曲における楽曲断片の検出手法を提案した．編曲された楽曲断片によって構成されているメドレー楽曲から，CRP 行列によって作成される累積値行列 Q を用いて，類似楽曲区間を検出し，Q_{max} を用いて定量的な値による楽曲断片検出を試みた．楽曲断片検出の精度評価において，提案した楽曲断片検出アルゴリズムが不適切であるもしくは，メロディ抽出が適切に抽出できていないことによる検出精度の低下が示唆された．今後は提案した楽曲断片候補の検出手法と Q_{max} を用いた楽曲断片の区間を決定する手法の妥当性についてまず検証する必要がある．必要であれば楽曲断片検出のアルゴリズム改良や，用いるメロディ抽出アルゴリズムの変更などによって精度の向上を図る．

謝辞　本研究を通じて，ご指導を賜りました寺井あすか准教授（公立はこだて未来大学）に深く感謝いたします．本研究は JSPS 科研費 16H01744，26280089 の助成を受けたものです．

参考文献

[1] Dannenberg, R. B. and Goto, M.:Music Structure Analysis from Acoustic Signals, In D. Havelock, Kuwano, S., Vorländer, M., editors, Handbook of Signal Processing in Acoustics, pp.477-482 (2011).
[2] Foote, J.: Visualizing Music and Audio using Self Similarity, In Proc. ACM International Conference on Multimedia, pp.77-80 (1999).
[3] Goto, M.: Active Music Listening Interfaces Based on Signal Processing, In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp..1441-1444 (2007).
[4] Kosugi, N., Sakurai, Y. and Morimoto, M.: SoundCompass: A Practical Query-by-Humming System, In Proc. ACM International Conference on Multimedia, pp.881-886 (2004).
[5] Marwan, N., Romano, M. C., Thiel, M. and Kurths, J.: Recurrence Plots for the Analysis of Complex Systems. Physics Reports, vol.438, No.5, pp237-329 (2007).
[6] Masuda, T., Yoshii, K., Goto, M. and Morishima, S.: Spotting a Query Phrase from Polyphonic Music Audio Signals based on Semi-supervised Nonnegative Matrix Factorization, International Society for Music Information Retrieval Conference (ISMIR), pp.227-232 (2014).
[7] Salamon, J. and Gomez, E.: Melody extraction from polyphonic music signals using pitch contour characteristics, IEEE Transactions on Audio, Speech, and Language Processing (TASLP), vol.20, No.6, pp.17591770 (2012).
[8] Serrà, J., Gómez, E. and Herrera P.: Transposing Chroma Representations to a Common Key, IEEE CS Conference on The Use of Symbols to Represent Music and Multimedia Objects, pp.45-48 (2008).
[9] Serrà, J., Gómez, E., Herrera P. and Serra, X.: ChromaBinary Similarity and Local Alignment Applied to Cover Song Identification, IEEE Trans. on Audio, Speech, and Language Processing (TASLP), Vol.16, No.6, pp.1138-1152 (2008).
[10] Serrà, J., Serra, X. and Andrzejak, R. G.: Cross Recurrence quantification, New Journal of Physics, Vol.11, No.9, pp.093017 (2009).
[11] Wang, A.: An Industrial Strength Audio Search Algorithm, In Proc. International Society for Music Information Retrieval Conference (ISMIR), pp.7-13 (2015).