

スペクトログラムの階層的クラスタリングを用いた タイムスパン・セグメンテーション抽出について

澤田 隼^{1,†1,a)} 竹川 佳成¹ 平田 圭二¹

受付日 2017年6月11日, 採録日 2017年12月8日

概要: 本稿では, Generative Theory of Tonal Music (GTTM) を音楽のスペクトログラムに直接適用して階層的クラスタリングによってタイムスパン・セグメンテーションを生成する新しい方法を提案する. まず初めに, スペクトログラムを時間軸方向に分割し, 周波数方向に縦長の矩形 (bin) をピッチイベントとして, スペクトログラムを一連の bin の集合として考える. bin のテクスチャの特徴は, グレーレベル同時生起行列 (Gray level co-occurrence matrix: GLCM) を使用して抽出され, テクスチャ特徴量の時系列データを生成する. テクスチャ特徴量による隣接 bin 間の類似度によってフレーズの近接度および変化量が計算される. 並列性および反復性などの大域的な構造は, 一連の bin の自己相似性行列 (Self-similarity matrix: SSM) によって検出される. 隣接する bin 間の境界の強さを表す時系列データが与えられ, 隣接する bin をボトムアップに反復的に併合していくことで, 最終的にタイムスパン・セグメンテーションに対応する系統樹を生成するアルゴリズムを開発する. Mozart の K.331 と K.550 を入力して実験を行った結果, 音高や調和などの音楽知識をほとんど考慮していないにもかかわらず, 有望な結果が得られた.

キーワード: generative theory of tonal music, time-span segmentation, グレーレベル同時生起行列, 自己相似性行列, 系統樹

On Extracting Time-span Segmentation Using Hierarchical Clustering of Spectrogram

SHUN SAWADA^{1,†1,a)} YOSHINARI TAKEGAWA¹ KEIJI HIRATA¹

Received: June 11, 2017, Accepted: December 8, 2017

Abstract: We propose a new method of applying Generative Theory of Tonal Music directly to a spectrogram of music to produce a time-span segmentation as hierarchical clustering. We first consider a vertically long rectangle in a spectrogram (bin) as a pitch event and a spectrogram as a sequence of bins. The texture feature of a bin is extracted using a gray level co-occurrence matrix to generate a sequence of the texture features. The proximity and change of phrases are calculated by the distance between the adjacent bins by their texture features. The global structures such as parallelism and repetition are detected by a self-similarity matrix of a sequence of bins. We develop an algorithm which is given a sequence of the boundary strength between adjacent bins, iteratively merges adjacent bins in the bottom-up manner, and finally generates a dendrogram, which corresponds to a time-span segmentation. We conducted an experiment with inputting Mozart's K.331 and K.550 and obtained promising results although the algorithm does not take into account almost any musical knowledge such as pitch and harmony.

Keywords: generative theory of tonal music, time-span segmentation, gray level co-occurrence matrix, self-similarity matrix, dendrogram

¹ 公立はこだて未来大学
Future University Hakodate, Hakodate, Hokkaido 041-8655, Japan

^{†1} 現在, 公立はこだて未来大学大学院
Presently with Intelligent Information Systems, Future University Hakodate

^{a)} g2116022@fun.ac.jp

1. はじめに

楽譜に書かれた楽曲の構造や意味を分析する手法として Generative Theory of Tonal Music (GTTM) がある [11]. これはグルーピング構造分析と拍節構造分析を経て, 人間

の認知過程をふまえた音楽の階層的な構造を抽出する分析手法である。グルーピング構造分析は楽曲全体を音楽的にまとまりのあるグループに分割し、グループの階層構造を抽出する分析であり、拍節構造分析は楽曲から各階層ごとに強拍と弱拍の位置を示す階層構造を抽出する分析である。各分析は構成規則 (well-formedness rules) と選好規則 (preference rules) の2種類の規則からなる。構成規則は満たすべき基本的な構造の特性を示す規則であり、選好規則は経験豊かな聴衆の聴取によって好ましい構造を示す規則である。GTTMのタイムスパン簡約は、グルーピング構造分析と拍節構造分析をもとに構造的に重要な音の選出を繰り返すことで旋律中の各音符の重要度を二分木 (タイムスパン木) で表すことができる。それは楽曲の構造の記述にとどまらず、楽曲の構造の操作を可能にするものであった。タイムスパン木は認知的リアリティを持つことがDibbenによって確認されており [3], 人間が音楽を聴取した際の認知過程をふまえた音楽的に信頼できる分析が可能となる。

しかし、GTTM分析を計算機上に実現しようとする多くの未解決の問題があることが広く認識されている [6], [8].

- (1) タイムスパン木を生成するために定められた、音楽の情報を抽出する選好規則の競合を解決するために優先順位を決定するという問題
- (2) 楽譜上に書かれた楽曲しか扱うことができず、音響信号を対象としていないという問題

(1) に関して、選好規則の音楽的な要素には、音高や時間間隔の差や、局所的な構造や大域的な構造の制約、和声や拍節の構造によって作られる境界などが含まれる。一般的に、競合する選好規則のどちらか一方に優先権を与えることを考えるが、選好規則の優先順位を制御する規則はいまだ提案されていない。(2) に関して、GTTMは本来楽譜に書かれた楽曲を分析するために提案された。しかし、実際の分析対象の楽曲は音響信号である。GTTMの適用範囲を音響信号に拡張することができれば、幅広いスタイルやジャンルなど膨大な楽曲に対して音楽的に信頼できる分析をすることができる。

しかし、GTTMを音楽音響信号に直接適用するにはGTTMの選好規則を音楽音響信号を対象とした規則に変換する問題を解決する必要がある。楽譜は拍節 (x 軸) と音高 (y 軸) の二次元座標系として考えることができるため、拍節を時間に、音高を周波数に変換することによって選好規則はスペクトログラムに適用可能である。しかし、スペクトログラムには多くの複雑な調波構造や不安定なパターンが含まれているため、メロディやコードの各音符を正確に認識して分離することは困難である。したがって、本来のGTTMの選好規則を音楽音響信号やスペクトログラムから抽出された音符に適用する方法では、高い楽譜認識の精度および音源分離の精度が要求される。

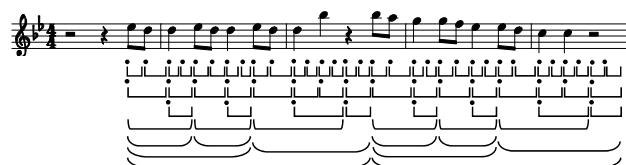


図 1 グルーピング構造と拍節構造を統合したタイムスパン・セグメンテーションの例

Fig. 1 Example of time span segmentation in which grouping structure and metrical structure are integrated.

本稿では、上記の問題を解決するために音楽音響信号に直接 GTTM を適用する新しい方法を提案する。音楽音響信号から抽出される代替の特徴量であるスペクトログラムのテクスチャ特徴に焦点を当てる。クロマグラムや MFCC などの音響特徴量の有効性を認めながらも、フィージビリティスタディとして、スペクトログラムのテクスチャ特徴量に基づいた新しい方法を検討し、タイムスパン・セグメンテーションを生成する。タイムスパン・セグメンテーションは Lerdahl らによって導入された基本的な音楽構造の1つで、グルーピング選好規則によるグルーピングの結果と、周期構造やリズムといった拍節構造の結果を統合した構造であり、タイムスパン簡約が行われる領域として定義されている。人間は旋律の特徴による上位のグループ境界の情報とリズムによる下位のグループ境界の情報を使って、人間の認知と整合する認知的リアリティのある境界を同定している。この2つの異なる境界の認知を統合する構造としてタイムスパン・セグメンテーションが導入された (図 1)。タイムスパン・セグメンテーションの例は、文献 [11] の図 6.5~6.6 (p.127) と図 6.8 (p.129) を参照されたい。また、タイムスパン・セグメンテーションの各セグメント内のヘッドを選択し、順序付けすることによってタイムスパン木が生成される。本稿では音高や和声の情報を扱うのではなく、スペクトログラムのテクスチャの特徴に着目するため、タイムスパン木ではなくタイムスパン・セグメンテーションを扱う。

2. 関連研究

まず、音響特徴量から音楽の境界や繰り返し構造などの音楽構造を抽出する典型的な従来の方法について述べる。Chen らは音楽音響信号をイントロや A メロ、サビといったセグメントに分解する方法を提案した [1]。Chen らクロマ特徴量に基づいた調和の情報や、MFCC に基づいた音色の情報を使用してセグメントにラベル付けを行った。Foote の自己相似性行列 (Self-similarity matrix: SSM) [4] のような音楽の構造を可視化すると同じ目的を果たす新しい表現 Score matrix が、音楽音響信号の調和と音色の2つの異なる側面を組み合わせるために導入された。そして、NMF によって Score matrix をセグメントのテンプレートと時系列のアクティベーションに分解する。Chen らの手法は主

に調和と音色といった様々な種類の音響情報に基づいて音楽構造が認識されるという観察から着想を得ている。

McFee らは複数のレベルの粒度で楽曲内の繰返し構造を効果的にエンコードする軽量の表現を提案した [12]。それらの手法はクロマグラムや MFCC の時系列データなどの音響特徴量からなる Recurrence matrix を生成することから始まる。ここでは大域的な繰返し構造を検出するための調和の特徴量と、局所的な整合性を検出するための音色の特徴量の 2 つの特徴量が使用される。次に、繰返し構造の発見を容易にするために、サンプル間の局所的な結合と長期的な結合が適切に調節され、最後に局所的な結合と大域的な結合のバランスをとり、繰返し構造をエンコードする Affinity matrix が得られる。セグメントのタイプの数を与えられると、ラブラシアングラフは Affinity matrix に適用され、楽曲構造を表したブロックを出力する。

Ullrich らは楽曲のセグメンテーションの問題に取り組み、サビや A メロなどの楽曲構造の境界を検出した [15]。Ullrich らはメルスケールされたスペクトログラムに人間がラベル付けしたコーパスを畳み込みニューラルネットワーク (CNN) が直接学習できるようにした。従来のセグメンテーションのアルゴリズムの多くは手作業で設計されており、精度を最適化するためにはチューニングが必要であるが、CNN を使用した教師あり学習は、その領域の知識なしの手作業の学習より優れている。CNN は、自身で音楽セグメンテーションに関連する特徴量を識別することができるため、コンピュータにとって有利である。

次に、ジャンルやムードで音楽を分類する従来研究を述べる。Costa らは音楽のジャンル分類のためにスペクトログラムのテクスチャ特徴量としてグレーレベル同時生起行列 (Gray level co-occurrence matrix: GLCM) と Local Binary Patterns (LBP) を採用した [2]。GLCM は直観的に、滑らかさ、粗さ、および規則性などの画像のテクスチャの特性の定量的な値を提供する [10] (詳しくは 3.1 節を参照されたい)。Haralick によって提案された 14 の特徴量のうち、Costa らは 7 つ (エントロピー、相関、均質性、3 次モーメント、最大尤度、コントラストおよびエネルギー) を使用し、分類器としてガウスカーネルの SVM を使用した。彼らは 2 つの異なる特徴抽出の戦略 (大域的と局所的) を検討した。前者は、スペクトログラム全体から特徴を抽出し、その後 1 つのジャンルに分類する。後者は、最初にスペクトログラムをいくつかの領域 (bin) に分割し、それぞれの bin をジャンルに分類し、すべての bin の部分的な分類を組み合わせる最終的なジャンルの決定を行う。その結果、前者よりもスペクトログラムを 5 つに分割した後者の方が優れており、高いパフォーマンスを達成した。

Nakashika らはテクスチャ特徴量として GLCM を使い、音楽のジャンル分類器に CNN を用いた [14]。Nakashika らはメルスケールされたスペクトログラムから異なる空間

関係 (距離と角度) を持つ複数の GLCM マップを提供した。いくつかの予備実験の後、Nakashika らは空間関係のパラメータの距離を 1 に設定し、角度を 0° , 45° , 90° および 135° に設定した。これらの異なる空間関係を有する GLCM マップのセットは、分類器の CNN への入力データを統合的に生成した。GLCM マップのセットは協調して局所的な音楽パターンを捕捉し、その結果単一の GLCM マップのみを使用した場合よりも優れていた。

3. スペクトログラムの階層的クラスタリング

グルーピング選好規則 (Grouping Preference Rule: GPR) は全部で 7 つあり、GPR2 (Proximity) と GPR3 (Change) は、それぞれピッチイベント間の時間軸方向の近接性 (Proximity) および、音高や音量などの変化 (Change) に基づいてグループの境界を形成する方法を規定した規則である。選好規則の各規則は付録 A.1 を参照されたい。ここでのグループの境界とは、グループとグループの境界を意味する。GTTM では構成規則によりグループに属さないピッチイベントや複数のグループに属するピッチイベントは存在しないよう規定されているため、グループの境界と境界の間がグループとなる。GPR4 (Intensification) は GPR2, 3 が満たす度合いが高いほど、GPR2, 3 によるグループの境界を形成する効果がより多く考慮されることを規定した規則である。たとえば、音高や音量などの変化が大きい場合は上位の階層でもグループの境界になる可能性が高いことを意味している。したがって、近接と変化の程度は実数で表現する。GPR2, GPR3 が局所的な視点で定義された境界であるのに対して、GPR5, GPR6 は大域的な視点で定義された境界である。GPR6 (Parallelism) は、繰返しの動機やフレーズは同じグルーピング構造になるという規則である。我々が興味があるのは、人間の聴覚が元々持っているゲシュタルト認知の能力だけで、音楽知識を用いずにスペクトログラムからどれだけ音楽的な構造分析ができるのかである。GPR1, 2, 3, 4, 6 は人間のゲシュタルト認知の能力を表したローレベルの性質であるのに対し、GPR5 (Symmetry) は比較的音楽的な要素が強いため現時点では考慮していない。また、GPR7 (Time-span and prolongational stability) はタイムスパン木が安定するようなグループ分割が望ましいという規則であり、本手法ではタイムスパン木が未定義のため考慮していない。

図 2 に本手法のスペクトログラムの階層的クラスタリングの概要を示す。初めに、入力された音楽音響信号を 1,024 サンプルの窓幅で 256 サンプルずつずらしながら短時間フーリエ変換し、スペクトログラムを 256 階調のグレースケールで描画する。Ullrich ら [15] や Nakashika ら [14] にならって、スペクトログラムの周波数軸はメルスケールを採用した。次に、スペクトログラムを時間軸方向にビートの位置で分割し、ビートの長さ分の短冊状のデータ (bin)

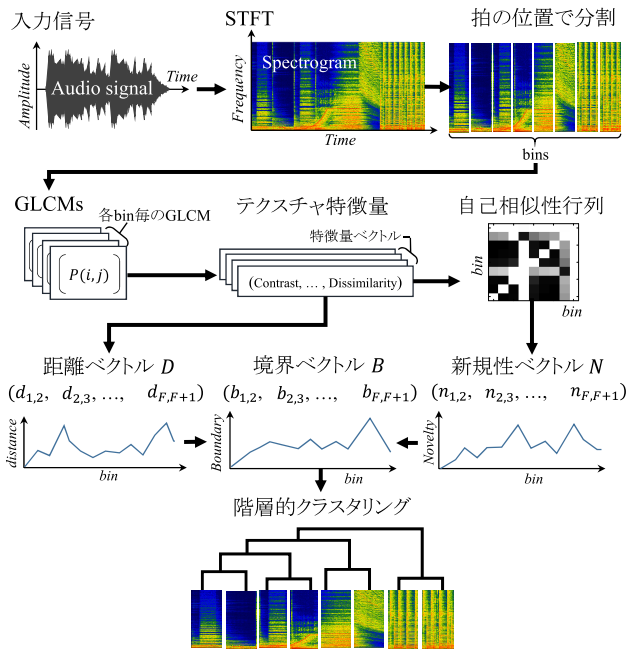


図 2 システム構成

Fig. 2 System configuration.

の集合として考える。ビート同期技術はMcFeeら[13]と同じ技術を採用した。Costaら[2]の研究ではスペクトログラムをいくつかのbinに分割する方法の方が、全体的に処理するよりも優れていることを示している。GTTMの最小単位はピッチイベントであるが、本研究ではこのbinが最小単位となる。これは単一のピッチイベントによるグループを形成するのは避けるというGPR1 (Alternative form)を満たしている。GTTMのグループには4分音符1拍より短い長さのグループは存在せず、それ以下の階層のグループは拍節構造によって分割されている。本手法では、ビート検出のためにlibrosaライブラリの`onset_detect`関数を使用した。なお、ビート検出の誤り箇所は手動で修正を行った。

次に各binごとにスペクトログラムのテクスチャ特徴量を抽出する。GPR2, 3が規定するピッチイベントの近接性や変化は、スペクトログラム上ではテクスチャのパターンとして表れる。パターン認識技術を使用し、スペクトログラム内の隣接するbin間の距離を計算し、これを近接および変化の尺度として使用する。詳しくは3.1節で述べる。その後、楽曲内の繰返し構造を抽出するためにこのテクスチャ特徴量を用いて自己相似性行列を計算する。これはGPR6の思想に基づいており、繰返し構造の始点と終点でグループの境界が強く引かれるような設計をした。詳しくは3.2節で述べる。最後に、隣接するbin間のテクスチャの特徴量の距離と、楽曲内の繰返し構造の情報を用いて時間軸方向に制約を持つ階層的クラスタリングを行う。GPR4はGPR2, 3で選択された効果が比較的顕著な場合はより大きなレベルのグループ境界が引かれるという規則

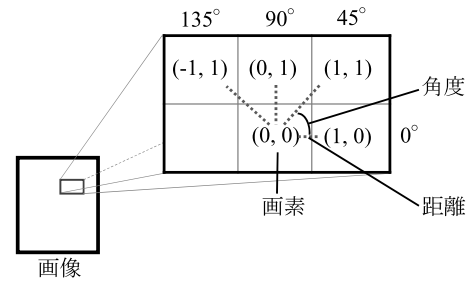


図 3 画素対の角度と距離

Fig. 3 Direction and distance of pixel pair.

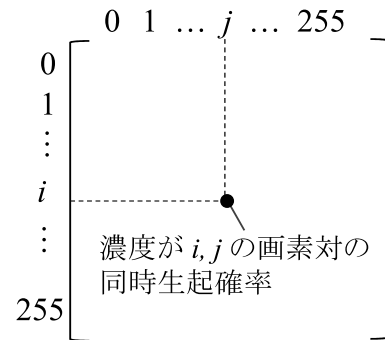


図 4 GLCM

Fig. 4 Gray level co-occurrence matrix.

であり、変化の大きい場合はより大きなレベルのグループ境界になるという意味で階層的クラスタリングはこれを満たす。テクスチャの変化が小さいものから併合されていき、テクスチャの変化が大きい場合は上位の境界として抽出される。制約付きの階層的クラスタリングについては3.3節で述べる。

3.1 グレーレベル同時生起行列とテクスチャ特徴量

図2の2段目では、各binからグレーレベル同時生起行列 (Gray level co-occurrence matrix: GLCM) [10]を用いてテクスチャ特徴量が抽出される。GLCMは画像中の隣接する画素対における同時生起する濃度の頻度を表す行列である(図4)。初めに、着目する画素の濃度と特定の空間関係に位置する隣接画素の濃度との共起を考える。次に、たとえば図3の4つの空間関係について、画素対の濃度が (i,j) であるとき、GLCMの (i,j) の位置の要素が1増加する。一般的にGLCMの各要素は、すべての要素の和が1.0になるように0.0から1.0の値に正規化される。この定義からも明らかのように、GLCMはパターンの平行移動に対して不変である。したがって、GLCMが周波数軸が対数スケールで描画された音楽音響信号のスペクトログラムに適用される場合、GLCMは平行に位置するフレーズに対して頑健である。GPR3は相対的な音高差によってグループの境界が形成される。また、GTTMでは音形やリズムが似ている部分は同じようなグループになることが期待されている。したがって、周波数方向に平行移動した「ドレ

ミ」と「ファソラ」は同じパターンと考え、似たような特徴量が抽出されることが期待されている。本手法の画素対の空間関係は Nakashika ら [14] と同様の4つの方向と距離を使用した。方向は 0° , 45° , 90° , 135° を扱い、画素対の距離は1ピクセルを採用した。

さらに、Haralick は画像の高次統計情報を表すコントラストや非類似性など、GLCM から二次特徴量を計算することによってテクスチャを分類する方法を提案した [10]。Haralick によって提案された14の二次特徴量のうち、コントラスト、非類似性、均質性、角二次モーメント (ASM) および相関の5つを採用した。コントラストや非類似性は、画素対の濃度差に関連した定義である。大きな濃度差を有する画素対の数が多いほど、コントラストと非類似性の値が高くなる。しかし、コントラストの値は指数関数的に増加するのに対して、非類似性の値は直線的に増加する。均質性は、GLCM の要素の分布が GLCM の対角成分に対してどの程度接近しているかを示す値である。これは、対角成分は画素対の濃度が (i, i) の頻度を表すため、濃度差が小さい画素が隣接する頻度が高い場合に均質性は大きい値を示す。対角成分から離れた要素（濃度差が大きい画素）の数が増加すると、均質性の値は指数関数的に減少する。ASM の値はテクスチャが整然としているときに高い値を示す。すべての画素が同じ値のときに最大値1.0をとる。相関はイメージ全体について、あるピクセルがその近傍とどのように相関しているのかを示している。これらの二次特徴量の数学的な定義については文献 [10] を参照されたい。

bin の GLCM が与えられると、上記の5つの二次特徴量が計算され、各特徴量の値の平均が0.0、分散が1.0に標準化される。最後に、標準化された値から、スペクトログラムの部分集合である bin のテクスチャ特徴量を表す5次元の特徴ベクトルを構築する。

3.2 境界ベクトル

距離ベクトル D は、一連の時系列データの特徴ベクトルから隣接する2つの特徴ベクトル間のユークリッド距離を計算することで生成される。たとえば図2の3段目において、 i 番目のビンと j 番目のビン間の距離は $d_{i,j}$ と表す。一連の bin の時系列の集合からタイムスパン・セグメンテーションを表す系統樹を作成するとき、最も類似した bin は基本的にボトムアップ方式でグルーピングされる。しかし、その際に考慮すべき点が2点ある。

- (i) タイムスパン・セグメンテーションの大域的な特性 (対称性や並列性)。
- (ii) 通常の系統樹を作成するアルゴリズムは隣接していない2つの bin を併合する。

(i) に関しては、新規性 (Novelty) [5] を導入した。新規性は、上述の特徴ベクトルから作成された自己相似性行

列の対角線に沿ってチェッカーボード・カーネル行列をずらしながらかけ合わせることで計算される。本手法では、チェッカーボード・カーネル行列のサイズは 2×2 を使用した。自己相似性行列は、繰返し構造や節などの比較的大きな構造の情報を反映するため、得られた新規性のピークは、その大域的な構造を考慮した音楽の構造境界の強度を意味する。本手法の新規性は図2の3段目にある特徴ベクトルのように、新規性ベクトル N の形で表される。次に拍節構造ベクトル M を導入する。拍節構造ベクトルは小節頭や強拍の位置といった拍節の階層的な情報を反映し、それらの位置で境界を強くするために境界ベクトルに重み付けを行うベクトルである。たとえば4拍子の場合その拍節構造ベクトル M は

$$M = (1.0, 0.25, 0.5, 0.25, 1.0, \dots, 0.25)$$

となる。最後に、隣接する bin 間の境界の総強度を表す境界ベクトル B は、距離ベクトル D と新規性ベクトル N と拍節構造ベクトル M の i 番目の要素を乗算することによって得られる。すなわち、 $b_{i,i+1} := d_{i,i+1} \times n_{i,i+1} \times m_{i,i+1}$ である。

(ii) に関しては、次の節でタイムスパン・セグメンテーションのための隣接する bin どうしのみを併合するように改良した新しい階層的クラスタリングのアルゴリズムについて述べる (図2の一番下の段)。

3.3 階層的クラスタリング

図5に、タイムスパン・セグメンテーションのための階層的クラスタリングのアルゴリズムを示す。図6には、bin がより大きな bin に併合される例を示す。

本手法では、2つの bin が併合されるたびに新たに併合

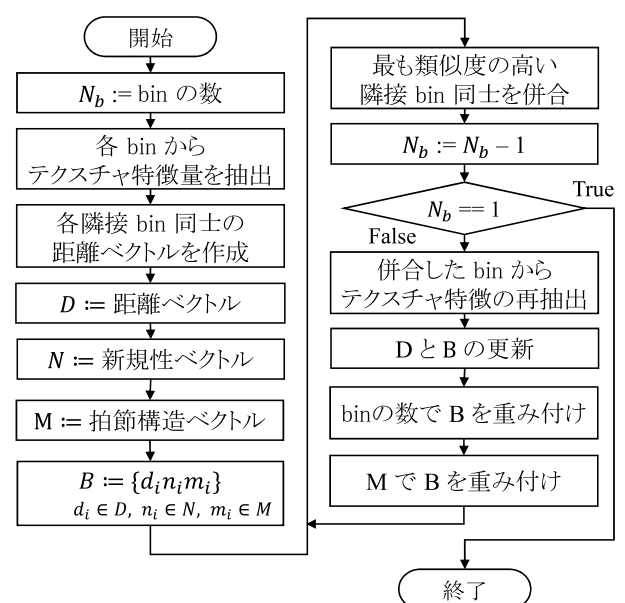


図5 階層的クラスタリングのアルゴリズム

Fig. 5 Hierarchical clustering algorithm.

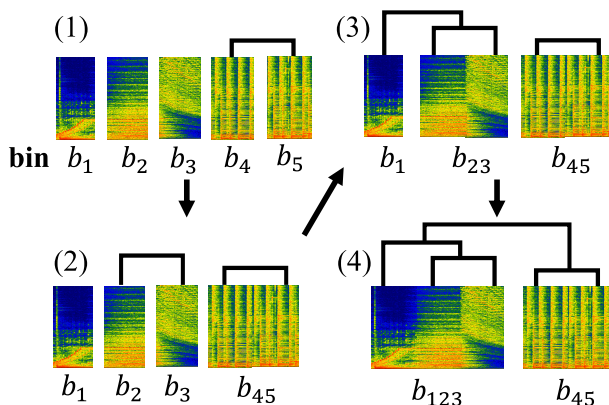


図 6 bin の併合の流れ
Fig. 6 Flow of merge of bin.

された bin の GLCM 特徴量が再計算され、距離ベクトルと境界ベクトルも再計算される。図 5 のアルゴリズムの最初の 7 つのステップは前の節で説明した初期化処理である。後半の 6 つのステップはすべての bin が単一の bin (元のスペクトログラム全体) に併合されるまで、距離ベクトル D および境界ベクトル B (新規性ベクトル N ではない) を更新しながら bin を反復的に併合するループ処理である。「bin の数で B を重み付け」のステップでは、併合された bin に含まれる単位 bin の数によって境界の強度が増大するように重み付けを行う。重み付け処理は併合された bin が大きいほど境界の強度が増大し、併合された bin が隣接する bin に併合されにくくなるという観察に基づく重み付けである。重み付け境界ベクトルの値を確認すると、最も類似した隣接する bin 間が最も弱い境界を有するため、より大きな bin に併合された。「M で B を重み付け」のステップでは、bin が併合されたことにより更新された距離ベクトルに、拍節構造ベクトルによって再度重み付けを行う (拍節構造ベクトルは更新されない)。拍節構造ベクトルによる重み付けは拍節構造の小節頭や強拍の位置で境界の強度が増大するという観察に基づく重み付けである。

bin が併合され、系統樹が作成される過程を図 6 に示す。たとえば、ステップ (1) では bin b_4 と b_5 が最も類似しているため、それらは b_{45} に併合される*1。 b_4 と b_5 を結ぶ線分の高さは境界線の強さを表す。線分が高いほど境界の強さは強くなる。

ステップ (2) では b_1 , b_2 , b_3 および b_{45} に対して距離ベクトル D と境界ベクトル B が最初に再計算される。次に、境界の強度を次に述べるように重み付けする。境界の強さ $b_{3,45}$ において、 b_3 は単一の bin からなり、 b_{45} は 2 つの bin からなるため、 $b_{3,45}$ は 2 ($= 1 \times 2$) で重み付けされ、 $b'_{3,45}$ ($= 2 \times b_{3,45}$) が得られる。その結果、境界の強度は増大され、 b_3 と b_{45} が併合されにくくなる。一方、境界の強さ $b_{2,3}$ の場合、 b_2 と b_3 の bin はどちらも単一の bin から

*1 $b_{i,i+1}$ は bin の b_i と b_{i+1} 間の境界の強さを表し、 $b_{i,i+1+i+2}$ は b_i と $b_{i+1+i+2}$ 間の境界の強さを表す。

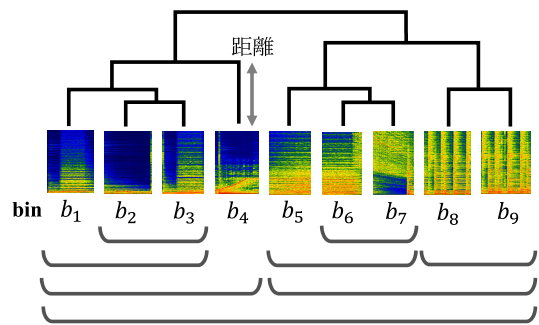


図 7 系統樹のタイムスパン・セグメンテーションへの変換
Fig. 7 Conversion of dendrogram to time span segmentation.

なるため、重み付けされた境界の強さ $b'_{2,3}$ は $b_{2,3}$ と同じままである。次に、拍節構造ベクトル M (1.0, 0.25, 0.5, 0.25) によって $b'_{1,2}$, $b'_{2,3}$, $b'_{3,45}$ に重み付けが行われる。 b_2 が小節の頭だとすると $b'_{1,2} = b_{1,2} \times 1.0$, $b'_{2,3} = b_{2,3} \times 0.25$, $b'_{3,45} = b_{3,45} \times 0.5$ が得られる。その後、 $b'_{2,3}$ と $b'_{3,45}$ が比較され、 $b'_{2,3}$ の方が $b'_{3,45}$ よりも弱いいため、 b_2 と b_3 が併合される。

ステップ (3) では b_1 , b_{23} および b_{45} に対して距離ベクトル D と境界ベクトル B が再計算される。次に、 b_{23} と b_{45} は両方とも 2 つの bin からなるため、境界の強さ $b_{23,45}$ は 4 ($= 2 \times 2$) で重み付けされ、 $b'_{23,45}$ ($= 4 \times b_{23,45}$) が得られる。一方、 b_1 は単一の bin からなり、 b_{23} は 2 つの bin からなるため、 $b_{1,23}$ は 2 で重み付けされ、 $b'_{1,23}$ ($= 2 \times b_{1,23}$) が得られる。それに加えて、拍節構造ベクトル M によって重み付けされ、 $b'_{1,23}$ ($= b'_{1,23} \times 1.0$) と、 $b'_{23,45}$ ($= b'_{23,45} \times 0.5$) が得られる。最後に $b'_{1,23}$ と $b'_{23,45}$ を比較し、 b_1 と b_{23} の bin が併合される。

3.4 タイムスパン・セグメンテーション

アルゴリズムがすべての bin を併合し終わると、bin を併合する過程を表す系統樹が得られる (図 7)。系統樹をタイムスパン・セグメンテーションに変換するのは容易である。得られた系統樹はボトムアップに解析され、系統樹における 2 つの bin を併合する点が見つかったと、これらの bin に対応するグループを内包する新しいグループが形成される。このようにして得られたタイムスパン・セグメンテーションは文献 [11], pp.37–39 に示されている 5 つの構成規則を満たすという意味で適切な定義である。同様の長さのグループ (たとえば 1 拍や 4 拍の長さ) は、ほぼ同じ持続時間として認識されるので、通常は同じ高さにプロットされる。

4. 実験結果

本手法を実証するために、Alfred Brendel が演奏する Schubert の Moments Musicaux の第三番 (Op.94-3) のテーマと、RWC Music Database [17] から、Mozart の piano sonata in A major (K.331) の第一楽章のテーマを

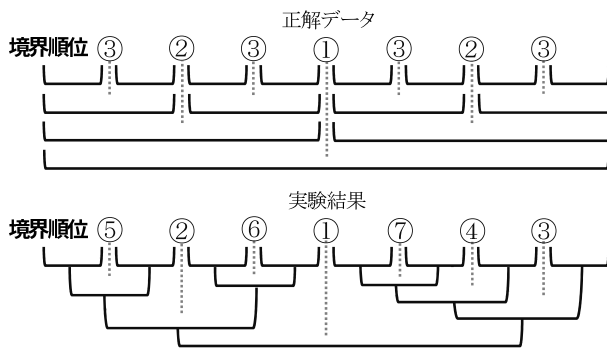


図 8 正解率の出し方

Fig. 8 How to calculate the accuracy rate.

表 1 本手法の F 値

Table 1 F-measure of this method.

対象楽曲	再現率	適合率	F 値
Op.94-3 (図 9)	0.55	1.0	0.71
K.331 (RWC, 図 10)	0.19	0.50	0.28
K.331 (Pires, 図 11)	0.27	0.70	0.39

使用した (RWC-MDB-C-2001 Nos. 26). それに加えて, Maria João Pires が演奏する K.331 を使用して, 同じ部分の比較を行った. ここで, 使用した Op.94-3 と K.331 はピアノ演奏によるホモフォニである. 正解データは元々の GTTM のルールを楽譜に適用した場合の結果を正解と見なし, GTTM の原本に書かれているものと, 浜中によって公開されている GTTM database [9] を使用した. これらの正解を図 9, 図 10, 図 11 の下部に示し, それぞれの F 値を表 1 に示す.

正解率の算出方法を図 8 を例に示す. 境界順位が高い順にトップダウンに再現率, 適合率を求める. 正解データの k 番目の境界集合を A_k , システム出力の k 番目の境界集合を B_k とすると, k 番目までの正解集合 $TP_k = \{A_1 \cup A_2 \cup \dots \cup A_k\} \cap \{B_1 \cup B_2 \cup \dots \cup B_k\}$ と書ける. 正解データの境界順位が N 番目までである場合, TP_k を 1 から N 番目まで順番に算出する. 再現率は以下となる.

$$\frac{\sum_{i=1}^N \sum_{k=1}^i |TP_k|}{\sum_{i=1}^N \sum_{k=1}^i |A_k|} \quad (1)$$

図 8 の例では再現率は 0.55, 適合率は 1.0, F 値は 0.71 となる.

4.1 Schubert の Moments Musicaux

図 9 に Alfred Brendel によって演奏された Op.94-3 の結果を示す. 本手法は正しく b_4 と b_5 間に最も強い境界を抽出し, b_2 と b_3 間と b_6 と b_7 間にも正しく次のレベルの境界を抽出した. SIGMUS114 で報告した方法では [16], 拍節構造ベクトル M を導入しておらず, この方法で同曲を分析すると拍節構造を取り入れることにより, b_6 と b_7 間の初期の併合を防ぐことができ, 精度が向上したことを示

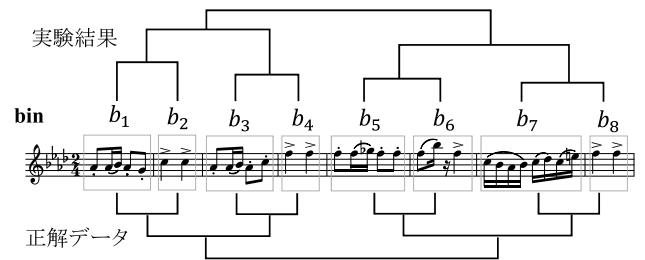


図 9 Moments Musicaux (Alfred Brendel) から得られた系統樹
Fig. 9 Dendrogram obtained from Moments Musicaux (Alfred Brendel).

している*2. 再現率は 0.55, 適合率は 1.0, F 値は 0.71 となった.

4.2 Mozart の Piano Sonata in A Major, K.331

図 10 に RWC Music Database の K.331 の結果を示す. 最も強い境界は b_8 と b_9 間であるにもかかわらず, システムは最も強い境界は b_5 と b_6 間であると出力した. b_1 と b_2 , b_3 と b_4 , b_9 と b_{10} , b_{11} と b_{12} のような最も低いレベルでのペアは, 初期段階で併合されるべきであった. これらの間違ったペアが系統樹の作成の初期段階で形成されたため, 間違ったペアの影響が上位のレベルに伝播されたと考えられる. 再現率は 0.19, 適合率は 0.50, F 値は 0.28 となった. 図 11 に Maria João Pires によって演奏された K.331 の結果を示す. RWC の結果とは対照的に, b_8 と b_9 間の最も強い境界が正しく検出され, 前述した最低レベルのペアのうち, 前半部分 (b_1 と b_2 や, b_3 と b_4) は正しく併合されたが, 後半部分 (b_9 から b_{16}) の系統樹の構成は正解から遠いものであった. 後半部分で正しい結果が得られなかった理由は図 12 と同様の初期段階の併合処理である. つまり, b_{12} と b_{13} が初めに併合されたため, b_{13} と b_{14} が併合されなかった. 再現率は 0.27, 適合率は 0.70, F 値は 0.39 となった.

5. 考察

最も低いレベルで起こっている併合の間違いの理由を考察する. K.331 (図 10) では, b_1 と b_2 間の新規性が高く, b_2 から b_4 が 0 であった. したがって, 初めに b_2 と b_3 が併合され, b_3 と b_4 間の境界の強度が更新され, 高くなる. その結果 b_4 と b_5 が併合される. 拍節構造ベクトルの導入により, 初期段階での b_8 と b_9 間の併合を未然に防ぐことができたが, 新規性が小さかったため, b_{10} と b_{11} 間の併合を防ぐことができなかった. その影響で b_{12} から b_{16} のグルーピングも正しく行われなかった.

これとは対照的に, Maria João Pires が演奏した K.331 の結果 (図 11) は良好であった. これは, 新規性が正しく

*2 拍節構造ベクトル M を組み込む前は b_6 と b_7 間が初期段階で併合されることにより, 最も強い境界が b_5 と b_6 間に抽出されていた.

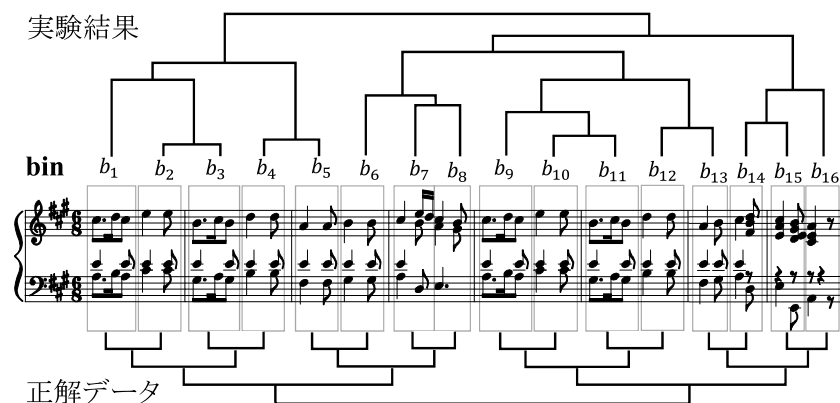


図 10 K.331 (RWC Music Database) から得られた系統樹

Fig. 10 Dendrogram obtained from K.331 (RWC Music Database).

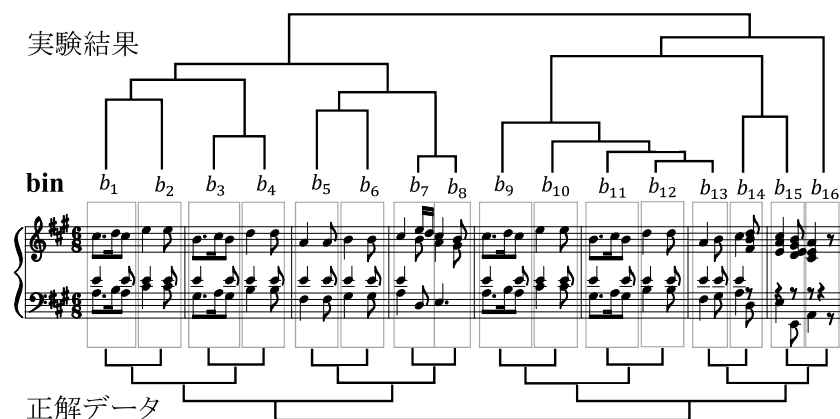


図 11 K.331 (Maria João Pires) から得られた系統樹

Fig. 11 Dendrogram obtained from K.331 (Maria João Pires).

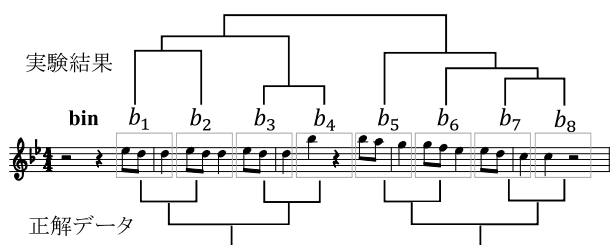


図 12 K.550 (RWC Music Database) から得られた系統樹

Fig. 12 Dendrogram obtained from K.550 (RWC Music Database).

計算されたからである．我々はチェッカーボード・カーネルのサイズが重要であると考えている．本手法における自己相似性行列 (SSM) のサイズは，K.331 では 16×16 であるため，大きなチェッカーボード・カーネルを使用することはできず，実際に用いたチェッカーボード・カーネルは 2×2 である．しかし，Foote [5] による本来の手法では，チェッカーボード・カーネルは 64×64 などの，より大きなサイズが使用されている．したがって，小規模な SSM に適した新規性の計算法を開発する必要がある．

また，K.331 (図 10) でのトップから 3 段目の併合での間違いについて， b_{1-5} と b_{6-8} ， b_{9-13} ， b_{14-16} の 4 つの bin が比較される際， b_{6-8} と b_{9-13} 間に強い境界が検出される

ことが期待されていたが，併合されてしまった．併合した単位 bin の数がそれぞれ (5, 3, 5, 3) となり，すべての境界が 15 ($= 5 \times 3$) で重み付けされるため，bin の大きさの重み付けの効果がなかった．これを解決するには GPR5 (Symmetry)，GPR7 (Time-span and prolongational stability) などの選好規則を考慮する必要や，音高やリズムなどの音楽情報を考慮する必要がある．

最後に演奏の表情付けによる分析結果の変化について述べる．Maria João Pires が演奏した K.331 は RWC の音源に比べて表情豊かなデータである．K.331 の結果 (図 11) の楽曲は b_8 と b_9 の間に最も強い境界が正しく検出されているが， b_7 から b_8 にかけて徐々に弱くゆっくり弾くような演奏をしており， b_8 のスペクトルのパワーは小さく， b_9 とのスペクトルのパワーの差は大きくなっていた．また， b_{14} で急に盛り上がり， b_{16} では急速に弱く弾くような演奏をしているため， b_{13} と b_{14} の間や b_{15} と b_{16} の間のスペクトルのパワーの差が大きくなっていた．このように，演奏者が楽譜には書かれていない音量の変化を加えると，音響信号のパワーが変化し，スペクトログラム上では濃度が変化する．また，音価を変化させた場合，たとえば bin が長くなると残響などの影響でテクスチャが変化し，次の bin との間にグループの境界が引かれる可能性が高くなる．し

かしこれらはゲシュタルト認知に基づくとは一概には間違いとはいえず、演奏家の解釈がタイムスパン・セグメンテーションとして表れていると考えることもできる。

6. 今後の展望

本手法はスペクトログラムを入力としているので、本来対象とはしていないがポリフォニの音響信号を入力することもできる。以下に Mozart の G Minor Symphony (K.550) のオープニングのテーマの分析結果を示す (RWC-MDB-C-2001 Nos. 2)。ここで、使用した K.550 は弦楽四重奏によるポリフォニである。

6.1 Mozart の G Minor Symphony, K.550

図 12 に RWC Music Database の K.550 の結果を上部に、GTTM Database の正解データを下部に示す。本手法は正しく b_4 と b_5 間に最も強い境界を抽出し、前半の分析結果が正しいことを示した。しかし、後半には分析誤りがあった。 b_6 を b_5 か b_{78} と併合する場合、アルゴリズムは境界の強度 $b_{5,6}$ と $b_{6,78}$ を比較し、 b_6 と b_{78} を併合するという誤った判断を下した。系統樹におけるノードの高さは bin を併合する順序を表している。まず、 b_3 と b_4 が併合され、その後 b_7 と b_8 が併合される。しかし、あくまでもこの結果は K.550 をモノフォニと見た場合の正解であり、ポリフォニに対するタイムスパン・セグメンテーションがどうあるべきかを検討する必要がある。たとえば Hamanaka らは楽譜に書かれた楽曲を対象として GTTM の規則をポリフォニへの拡張を行った [7]。これはポリフォニからホモフォニにアレンジする作曲家の過程を観察し、タイムスパン分析のための新たなルールを実装することにより、ポリフォニのタイムスパン木を半自動的に獲得するものであり、ポリフォニに対するタイムスパン木の構造を定義している。

7. おわりに

本稿では、GTTM を音楽のスペクトログラムに直接適用して、タイムスパン・セグメンテーションを生成する新しい方法を提案した。音楽音響信号からスペクトログラムのテクスチャの特徴のみを使用してタイムスパン・セグメンテーションを抽出しようとする試みによる、図 9, 図 10, 図 11, 図 12 に示す結果は、我々が期待していたよりも有望であった。たとえば、対象範囲の拡大にともない、サンプル数の増加による分類精度の向上や、GTTM の認知的リアリティの保証が見込まれる。スペクトログラムを音響的知識をほとんど用いずに処理をしたにもかかわらず、GTTM が主張したデータ構造が得られた。これは GTTM が認知的な根拠を補強する要素になりうると考えている。この結果は、音楽における階層的クラスタリングが音楽特有の認知機能ではなく、人間が他のメディアを理解するた

めに使用している一般的な認知機能の 1 つであることを示唆している。本手法の精度を向上させるには、音高や調和、リズムなどの音楽情報が役立つと考えられる。

今後の課題は 3 点ある。

- 音高や調和、リズムなどの音楽情報を考慮した境界ベクトルの生成
- 未実装のグルーピング選好規則である GPR5 (Symmetry) や他の選好規則を実装した階層的クラスタリングのためのアルゴリズムの開発
- 楽曲数を増やした実験とその結果に対する定量的な評価

現在実装されている新規性を用いた選好規則は不十分であり、小規模な SSM に適した新規性の計算法を開発し、階層的な繰返し構造を抽出する必要がある。

謝辞 研究を通じて議論をさせていただいた寺井あすか先生 (公立はこだて未来大学) に感謝いたします。本研究は JSPS 科研費 16H01744 の助成を受けたものです。

参考文献

- [1] Chen, R. and Li, M.: Music Structural Segmentation By Combining Harmonic and Timbral Information, *Proc. ISMIR*, pp.477–482 (2011).
- [2] Costa, Y.M.G., Oliveira, L.S., Koerich, A.L. and Gouyon, F.: Comparing textural features for music genre classification, *Proc. 2012 International Joint Conference on Neural Networks*, pp.1867–1872 (2012).
- [3] Dikken, N.: Cognitive Reality of Hierarchic Structure in Tonal and Atonal Music, *Music Perception: An Interdisciplinary Journal*, Vol.12, No.1, pp.1–25 (1994).
- [4] Foote, J.: Visualizing Music and Audio using Self Similarity, *Proc. 7th ACM Int'l Conf. Multimedia*, pp.77–80 (1999).
- [5] Foote, J.: Automatic audio segmentation using a measure of audio novelty, *Proc. IEEE International Conference on Multimedia and Expo*, Vol.1, pp.452–455 (2000).
- [6] Hamanaka, M., Hirata, K. and Tojo, S.: Implementing “A Generative Theory of Tonal Music”, *Journal of New Music Research*, Vol.35, No.4, pp.249–277 (2007).
- [7] Hamanaka, M., Hirata, K. and Tojo, S.: Toward Developing a Polyphonic Music Time-Span Tree Analyzer, *Mathematics and Computation in Music 2013 (MCM2013)* (June 2013).
- [8] Hamanaka, M., Hirata, K. and Tojo, S.: Implementing Methods for Analysing Music Based on Lerdahl and Jackendoff's Generative Theory of Tonal Music, *Computational Music Analysis*, Meredith, D. (Ed.), Chapter 9, pp.221–249, Springer (2016).
- [9] 浜中雅俊: Interactive GTTM Analyzer/GTTM Database Download Page, 入手先 (<http://gttm.jp/gttm/ja/database/>) (2017).
- [10] Haralick, M.R.: Statistical and structural approaches to texture, *Proc. IEEE*, Vol.67, No.5, pp.786–804 (1979).
- [11] Lerdahl, F. and Jackendoff, R.: *A Generative Theory of Tonal Music*, The MIT Press (1983).
- [12] McFee, B. and Ellis, P.W.D.: Analyzing Song Structure with Spectral Clustering, *Proc. ISMIR*, pp.405–410 (2014).
- [13] McFee, B. and Ellis, P.W.D.: Learning to Segment

Songs with Ordinal Linear Discriminant Analysis, *Proc. ICASSP* (2014).

- [14] Nakashika, T., Garcia, C. and Takiguchi, T.: Local-feature-map Integration Using Convolutional Neural Networks for Music Genre Classification, *Proc. Interspeech*, pp.1752–1755, ISCA (2012).
- [15] Ullrich, K., Schlüter, J. and Grill, T.: Boundary Detection in Music Structure Analysis using Convolutional Neural Networks, *Proc. ISMIR*, pp.417–422 (2014).
- [16] 澤田 隼, 竹川佳成, 平田圭二: スペクトログラムの階層的クラスタリングを用いたグルーピング構造分析について, 情報処理学会研究報告音楽情報科学, Vol.2017-MUS-114, No.7, pp.1–8 (2017).
- [17] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases, *Proc. ISMIR*, pp.287–288 (2002).

付 録

A.1 Grouping Preference Rule (GPR)

GPR1 (Alternative form) 1つのピッチイベントだけでグループを形成するのは強くさける。

GPR2 (Proximity), GPR3 (Change) 連続した4つのピッチイベントをそれぞれ n_1 , n_2 , n_3 , n_4 とするとき, n_2 , n_3 の間で他のピッチイベントの間よりも以下の条件が成り立つときグループの境界と認識される。

GPR2a 休符が長い。

GPR2b オンセット時間の間隔が長い。

GPR3a 音程が大きい。

GPR3b 音量の変化が大きい。

GPR3c アーティキュレーションの変化が大きい。

GPR3d 音長の変化が大きい。

GPR4 (Intensification) GPR2, GPR3 で選択された効果が比較的顕著な場合は, より大きなレベルのグループ境界が引かれる。

GPR5 (Symmetry) グループの分割の長さが等しい2つの部分からなるようなグルーピングが望ましい。

GPR6 (Parallelism) グループ間で並行しているグループは並行性のあるグルーピングが望ましい。

GPR7 (Time-span and prolongational stability) タイムスパン木が安定するようなグループが望ましい。



澤田 隼 (学生会員)

に関する研究に従事。

2016 年公立はこだて未来大学システム情報科学部複雑系知能学科卒業。同年同大学大学院博士前期課程システム情報科学研究科に進学。現在, 博士前期課程 2 年。2016 年度音楽情報科学研究会学生奨励賞受賞。音楽情報処理



竹川 佳成 (正会員)

2007 年大阪大学大学院情報科学研究科博士課程修了。同年より神戸大学自然科学系先端融合研究環境重点研究部助教。2012 年より公立はこだて未来大学システム情報科学部助教。2014 年より公立はこだて未来大学システム情報科学部准教授, 現在に至る。博士 (情報科学)。ヒューマンコンピュータインタラクション, エンタテインメントコンピューティング, 音楽情報科学の研究に従事。



平田 圭二 (正会員)

1987 年東京大学大学院工学研究科情報工学専門課程博士課程修了。同年 NTT 基礎研究所入社。1990~1993 年 (財) 新世代コンピュータ技術開発機構 (ICOT) に出向。2011 年より公立はこだて未来大学教授, 現在に至る。博士 (工学)。1993~1995 年情報処理学会音楽情報科学研究会初代主査, 2010~2015 年デジタルプラクティス誌編集委員長。音楽情報処理, 実時間デマンド型公共交通 SAV, うつ病家族看護者の ICT 支援研究に従事。本会フェロー。