類似楽曲検索用オーディオ指紋の提案と評価

青島 大河[1,a)]　山森 一人[2,b)]　井上 健太郎[2,c)]　相川 勝[3,d)]

概要：音響信号の持つ単純な音響的特徴のみに基づく従来のオーディオ指紋は、近年需要が高まっている類似楽曲検索には向いていない。類似楽曲検索のために、音楽の観点に基づいた特徴をオーディオ指紋に用いる方法が考えられる。本稿では、音楽の 3 つの基本要素であるリズム、メロディ、ハーモニーに基づいた、類似楽曲検索のためのオーディオ指紋を提案する。実際の楽曲を用いた楽曲間距離の導出実験、および第三者による官能テストにより提案手法の妥当性を検証する。

キーワード：音楽情報検索, 類似楽曲検索, 短時間フーリエ変換

A Novel Audio Fingerprinting Method for Similar Music Retrieval

Taiga Aoshima[1,a)]　Kunihito Yamamori[2,b)]　Kentaro Inoue[2,c)]　Masaru Aikawa[3,d)]

Abstract: Conventional audio fingerprinting methods for fast music retrieval are based on the acoustic features only, and they are not suitable for similar music retrieval. In this paper, we propose an audio fingerprinting method for similar music retrieval focusing on the three fundamental music elements: rhythm, melody and harmony. We calculate the distances between actual music using the proposed method, then evaluate the validity of the proposed method.

Keywords: Music Information Retrieval, Similar Music Retrieval, Short-Time Fourier Transform

1. はじめに

近年、ブロードバンドサービスの普及、オンライン楽曲配信サービスの一般化などにより、多くの人がそうしたサービスを利用して気に入った楽曲を楽しんだり、新しい楽曲を探したりできるようになった。しかし、楽曲配信サービスの保有するデータベースが肥大化し、膨大な数の楽曲が保存されるようになると、その中から新たに気に入る楽曲を人手で探すのは大変な作業となる。そのため、コンピュータを用いた自動楽曲検索の研究が盛んに行われている [1]。

楽曲検索手法の中で最も一般的な手法は、キーワードによる楽曲検索である。検索に用いるキーワードとしては、楽曲名、アーティスト名、ジャンル、年代、歌詞などの情報が挙げられる。キーワードによる楽曲検索を用いた配信サービスは、YouTube [2] や Google Play Music [3] をはじめとして数多く存在する。しかし、楽曲にキーワードが付与されていなかったり、付与されていても曖昧であったりすると、キーワードを用いた楽曲検索ではユーザの求める楽曲が検索できない場合がある。

ユーザが気に入る可能性のある楽曲は、ユーザが現在気に入っている楽曲と類似している可能性が高いと考えられる。つまり、ユーザが所有している好みの楽曲と類似した楽曲を検索することにより、ユーザの求める楽曲が検索できると考えられる。多くの研究者が、コンピュータを用いた効率的な類似楽曲検索の研究を行っている [4–8]。しか

[1] 宮崎大学 工学研究科
Graduate Engineering, University of Miyazaki, Japan
[2] 宮崎大学 工学教育研究部
Faculty of Engineering, University of Miyazaki, Japan
[3] 宮崎大学 工学部 教育研究支援技術センター
Technical Center, Faculty of Engineering, University of Miyazaki, Japan
[a)] aoshima@taurus.cs.miyazaki-u.ac.jp
[b)] yamamori@cs.miyazaki-u.ac.jp
[c)] inoue@cs.miyazaki-u.ac.jp
[d)] aikawa@cs.miyazaki-u.ac.jp

し、ユーザの感じる楽曲の類似性は曖昧であり、類似楽曲を安定して検索できる手法は確立されていない。

楽曲検索の従来手法として、オーディオ指紋 [9–11] を用いた手法がある。オーディオ指紋とは、Haitsma と Kalker が 2002 年に提案した手法であり、楽曲の音響的特徴から生成される比較的コンパクトなバイナリデータ（オーディオ指紋）を用いて楽曲検索を行う。オーディオ指紋を用いることにより、膨大な楽曲データベースからの楽曲検索を高速に行うことができる。従来のオーディオ指紋は、同一楽曲を検索する用途としては有効であるが、同一でない、類似した楽曲を検索するのには有効でない。なぜなら、従来のオーディオ指紋は音響信号の持つ単純な音響的特徴のみに基づいており、音楽の観点からみた特徴を考慮していないためである。

類似楽曲検索手法の提案にあたり、音楽の 3 つの基本要素に着目した。音楽の 3 つの基本要素とは、リズム、メロディ、ハーモニーを指す [12,13]。2 つの楽曲が類似している場合、それらの楽曲に含まれる音楽の基本要素も類似していることが想定できる。このことから、音楽の基本要素の特徴を用いることにより、類似した楽曲を検索できる可能性を指摘できる。

本稿では、音楽の 3 つの基本要素であるリズム、メロディ、ハーモニーに基づいた、類似楽曲検索のための新たなオーディオ指紋を提案する。提案手法では、楽曲を短時間フーリエ変換（STFT）を用いて多くのパワースペクトルに変換し、各パワースペクトルから抽出した音楽の基本要素の特徴から 32 ビットのバイナリデータを生成し、それらを時系列順に連結してオーディオ指紋を生成する。生成したオーディオ指紋を用いて楽曲間距離を算出することにより、楽曲間の類似度を導出する。

本稿の構成は以下の通りである。第 2 章では、従来の Haitsma と Kalker のオーディオ指紋について説明する。第 3 章では、提案するオーディオ指紋について説明する。第 4 章では、提案手法を用いた楽曲の類似度導出の実験を行い、提案手法の評価を行う。第 5 章は本稿のまとめである。

2. 従来のオーディオ指紋

2.1 オーディオ指紋の概要

オーディオ指紋 [9] は、楽曲の音響的特徴やヒトの知覚的特性に基づき、元の音楽データをコンパクトなバイナリデータに変換したものである。オーディオ指紋には、MD5 等のハッシュアルゴリズムと異なり、入力となる音響信号データのノイズや劣化の有無にかかわらず、同一楽曲からは類似したバイナリデータが生成されることが特長として挙げられる。入力となる音響信号データと比較してデータサイズが小さいため、楽曲検索を高速に行うことができる。

1 つのオーディオ指紋は、複数のサブ指紋から構成される。1 つのサブ指紋は、STFT により楽曲片から変換された 1 つのパワースペクトルから生成される。各パワースペクトルから生成されたサブ指紋を時系列順に連結することにより、楽曲 1 曲分のオーディオ指紋が生成される。Haitsma と Kalker によるオーディオ指紋生成法は、3 つのステップからなる。図 1 に、Haitsma と Kalker によるオーディオ指紋生成の流れを示す。

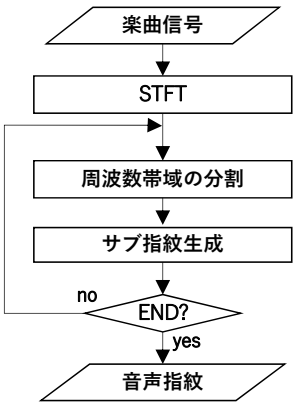

図 1 従来のオーディオ指紋生成手順.

ステップ 1 では、STFT を用いて楽曲片からパワースペクトルを導出する。ステップ 2 では、導出したパワースペクトルを 33 個の周波数帯域に分割する。ステップ 3 では、分割した 33 個の周波数帯域のエネルギーから、サブ指紋を生成する。

すべてのサブ指紋の生成後、それらを時系列順に連結し、1 つのオーディオ指紋を生成する。

2.2 パワースペクトルの導出

Haitsma と Kalker によるオーディオ指紋導出のステップ 1 では、サブ指紋を生成するために楽曲信号を STFT を用いてパワースペクトルに変換する。

まず、入力となる楽曲信号をオーバーラップした楽曲片に分割する。分割した楽曲片 1 つ当たりの長さは 0.37 秒であり、オーバーラップの長さは楽曲片の長さの 31/32 である。楽曲片 1 つにつきサブ指紋が 1 つ生成されるため、楽曲の演奏時間 11.6 ミリ秒（楽曲片 1 つの長さの 1/32）につき 1 つのサブ指紋が生成されることになる。

続いて、分割された各楽曲片を FFT を用いてパワースペクトルに変換する。FFT の際、各楽曲片の信号にはあらかじめハニング窓関数を掛け合わせる。

2.3 周波数帯域の分割

Haitsma と Kalker によるオーディオ指紋導出のステップ 2 では、ステップ 1 で導出したパワースペクトルを、周波数 300 Hz から 2,000 Hz までのオーバーラップしない 33 個の周波数帯域へ分割する。この範囲は、ヒトの聴覚（Human Auditory System：HAS）において最も聞き取りやすい音域である [9]。

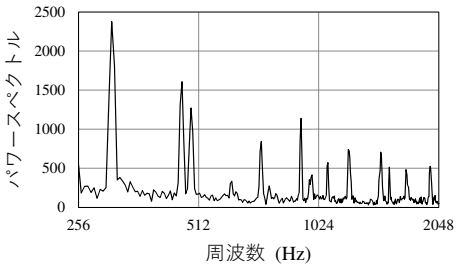

図 2 パワースペクトルの例.

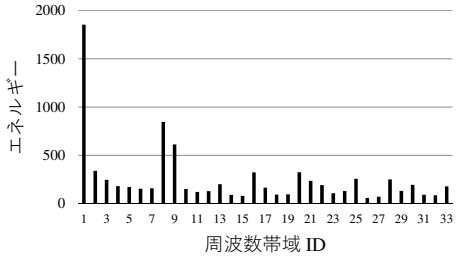

図 3 分割された周波数帯域の例.

図 2 にパワースペクトルの例を、図 3 に当該パワースペクトルを 33 個の周波数帯域に分割した結果を示す。

2.4 サブ指紋の生成

Haitsma と Kalker のオーディオ指紋導出のステップ 3 では、分割した周波数帯域のエネルギーから、サブ指紋を生成する。

t 番目のパワースペクトルの n 番目の周波数帯域のエネルギーを $E(t,n)$ とするとき、t 番目のパワースペクトルから生成するサブ指紋の、n 番目のビット $F(t,n)$ は、式 (1) で求められる。

$$F(t,n) = \begin{cases} 1, & if \quad ED(t,n) > 0, \\ 0, & if \quad ED(t,n) \leq 0. \end{cases} \tag{1}$$

$$\begin{aligned} ED(t,n) \;=\; & E(t,n) - E(t,n+1) \\ & - (E(t-1,n) - E(t-1,n+1)) . \end{aligned}$$

サブ指紋の各ビットは、ある時刻 t における隣接した 2 つの周波数帯域 $n, n+1$ のエネルギー差と、時刻 $(t-1)$ における当該周波数帯域のエネルギー差から算出される。

STFT によって得られたすべてのパワースペクトルからサブ指紋を生成したのち、そのサブ指紋すべてを時系列順に連結して 1 つのオーディオ指紋を生成する。

2.5 オーディオ指紋の比較

オーディオ指紋の比較には、1 ビット当たりの平均誤り率（Bit Error Rate：BER）を使用する。オーディオ指紋の比較は、サブ指紋ブロック単位で行う。サブ指紋ブロックとは、256 個のサブ指紋の時系列順の連結である。サブ指紋ブロック間の BER を式 (2) に示す。

$$BER(A,B) = \frac{\sum_{t=1}^{256} \sum_{n=1}^{32} [F_A(t,n) \oplus F_B(t,n)]}{32 \times 256}, \tag{2}$$

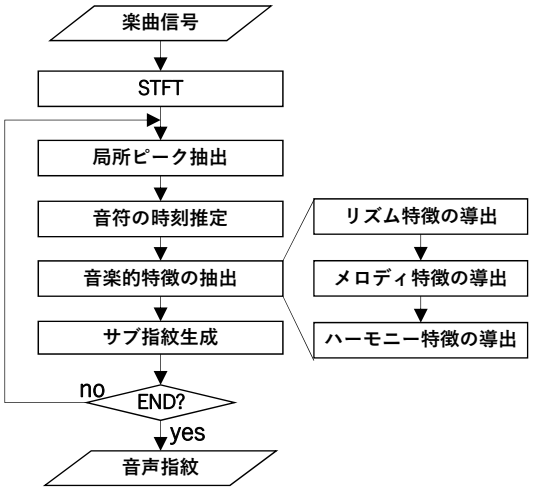

図 4 提案手法のオーディオ指紋生成手順.

ここで $F_A(t,n)$ は、曲 A から生成されたオーディオ指紋の、あるサブ指紋ブロックにおける t 番目のサブ指紋の n 番目のビットであり、$\oplus$ は排他的論理和（XOR）である。

3. 提案するオーディオ指紋

3.1 提案手法の概要

本稿では、音楽の 3 つの基本要素に焦点を当てたオーディオ指紋を提案する。提案するオーディオ指紋生成法は、5 つのステップからなる。図 4 に、提案するオーディオ指紋の生成手順を示す。

ステップ 1 では、STFT を用いて楽曲片からパワースペクトルを導出する。ステップ 2 では、導出したパワースペクトルから局所ピークとその周波数を推定する。推定した局所ピークは、ハーモニーの特徴の導出に用いられる。ステップ 3 では、推定した局所ピークから音の発生時刻と周波数を推定する。推定した音の発生時刻の情報は、リズムとメロディの特徴の導出に用いられる。ステップ 4 では、推定した局所ピークと音の発生時刻の情報から、3 つの音楽的特徴を抽出する。ステップ 5 では、抽出した 3 つの音楽的特徴をバイナリデータに符号化し、これらを連結することでサブ指紋を生成する。

すべてのサブ指紋の生成後、それらを時系列順に連結し、1 つのオーディオ指紋を生成する。

3.2 パワースペクトルの導出

提案手法のステップ 1 では、楽曲信号を STFT を用いてパワースペクトルに変換する。提案手法ではフーリエ変換の際、楽曲片の信号にはあらかじめハミング窓を掛け合わせる。

3.3 局所ピークの推定

提案手法のステップ 2 では、導出したパワースペクトルから局所ピークとその周波数を推定する。楽曲片のパワー

スペクトル上の局所ピークが持つ情報を用いて、音楽的な特徴を抽出する。パワースペクトル上の局所ピークの周波数は、楽曲片に含まれる音の音高に等しい。また、楽曲片に含まれる音の音高から、ハーモニーの特徴を導出できる。どの楽曲片にパワースペクトルの局所ピークが現れるかという情報から、音の発生時刻が推定できる。また、音の発生時刻と音高から、リズムとメロディの特徴を導出できる。

ある周波数帯域 n におけるパワースペクトルのエネルギーを $E(n)$ とする。このとき、$E(n)$ が式 (3) を満たす場合、$E(n)$ は局所ピークであると定義する。このステップでは、式 (5) を満たす局所ピークのみを抽出する。

$$E(n-1) \leq E(n) \geq E(n+1). \tag{3}$$

$$Note\,(f(n)) = \mathrm{round}\left(69 + 12\log_2\left(\frac{f(n)}{440}\right)\right). \tag{4}$$

$$54 \leq Note\,(f(n)) \leq 84. \tag{5}$$

式 (4) において、$f(n)$ は周波数帯域 n の周波数を示し、$Note(f(n))$ は周波数 $f(n)$ に対応するノートナンバーを示す。ノートナンバーとは、MIDI フォーマットにおいて音の音高を表す値である。式 (4) の定数 69 は、周波数 440 Hz に対応するノートナンバーである。

式 (5) における 54 と 84 はそれぞれ、周波数 185.0 Hz、1,046.5 Hz に対応するノートナンバーである。この範囲は、日本の楽曲に含まれる音の音高が G2（196.0 Hz）から G4（784.0 Hz）であること [18]、および HAS において最も聞き取りやすい音域が 300 Hz から 2,000 Hz であること [9] から定めた。

$E(n)$ を局所ピークと判定した後、Quadratically Interpolated FFT（QIFFT）法 [19–21] を使用し、$E(n)$、$E(n-1)$、$E(n+1)$ から尤もらしい局所ピークの周波数とエネルギーを推定する。QIFFT 法の概要を図 5 に示す。$E(n-1)$、$E(n)$、$E(n+1)$ は離散的な値であり、最尤局所ピークは $E(n-1)$ から $E(n)$ の間、もしくは $E(n)$ から $E(n+1)$ の間に存在すると考えることができる。QIFFT 法は上記 3 つの点を通る 2 次曲線を導出し、二次曲線の頂点を最尤局所ピークとみなすことにより、パワースペクトルの最尤局所ピークの周波数とエネルギーを推定する手法である。

3.4 音の発生時刻の推定

提案手法のステップ 3 では、推定した局所ピークから音の発生時刻と周波数を推定する。これらの情報は、リズムとメロディの特徴の導出に用いる。

楽曲片 t が周波数 f に局所ピークを有するとき、その局所ピークを $P(t, f)$ と表す。$P(t, f)$ が楽曲片 t に存在する一方、楽曲片 $(t-1)$ に存在しない場合、楽曲片 t で周波数 f（ノートナンバー $Note(f)$）の音が発生したとする。

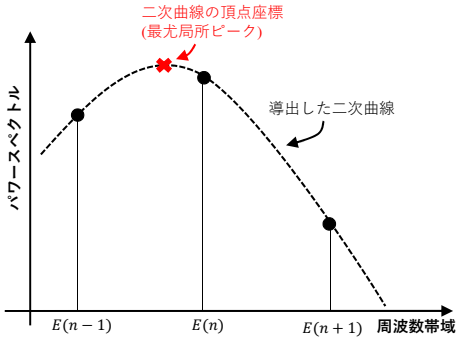

図 5　QIFFT 法による最尤局所ピークの推定.

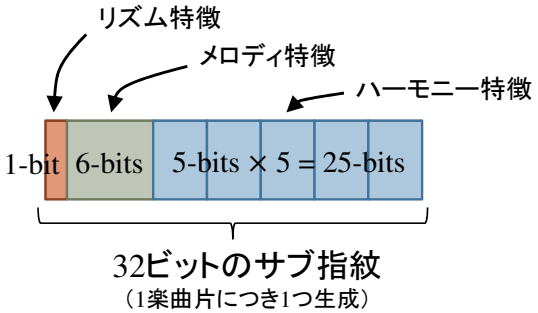

図 6　サブ指紋の概要.

3.5 サブ指紋の生成

提案手法のステップ 4 では、以下の節で説明する音楽の 3 つの基本要素の特徴を楽曲片の局所ピークから抽出し、それらの特徴をバイナリデータに符号化、3 つのバイナリデータをリズム、メロディ、ハーモニーの順番に連結することで 1 つのサブ指紋を生成する。STFT を用いて得られたパワースペクトル 1 つから、1 つのサブ指紋が生成される。図 6 に、サブ指紋の概要を示す。

1 つのサブ指紋は 32 ビットのバイナリデータである。提案手法では図 6 に示すように、リズム、メロディ、ハーモニーの特徴は、それぞれ 1 ビット、6 ビット、25 ビットのバイナリデータで表される。

3.5.1 リズム特徴

リズムの観点から、類似した楽曲は類似した音の発生分布を持っていると考えられる。このことから本稿では、音の発生時刻をリズム特徴として用いる。

サブ指紋のリズム特徴は 1 ビットのバイナリデータで表される。楽曲片で音の発生がある場合、その楽曲片から生成されるサブ指紋のリズム特徴には 1 がセットされ、ない場合は 0 がセットされる。

3.5.2 メロディ特徴

メロディとは、さまざまな音高の音の連結である。メロディの観点から、類似した楽曲は類似した音高の変化を持っていると考えられる。このことから本稿では、隣接する音高の変化をメロディ特徴として用いる。音高はノートナンバーで表されており、式 (5) で示した通り 54 から 84 の範囲に制限されているため、音高の変化は-30 から 30 の

値をとる。したがって、サブ指紋のメロディ特徴を表すために 6 ビットを与える。

3.5.3 ハーモニー特徴

ハーモニーの観点から、類似した楽曲は類似したコードを持っていると考えられる。コードとは、同時に存在する 3 つ以上の音の集合である。現代の典型的な楽曲は主に、トライアド（3 種類の音高）、セブンスコード（4 種類の音高）、ナインスコード（5 種類の音高）の 3 種類のコードから構成されている。楽曲片から音の音高を抽出することができれば、その楽曲片に含まれるコードが推定できる。コードを推定することができれば、ハーモニー特徴を導出できる。コードの中で最も構成音が多いのは、5 音で構成されるナインスコードであるため、本稿ではハーモニー特徴として最大 5 つの音の音高を用いる。

1 つの音の音高は 5 ビットのバイナリデータで表される。音高のノートナンバーは式 (5) より 54 から 84 の範囲であるため、ノートナンバーから 53 を引いた数値（1 から 31 の範囲）をハーモニー特徴の音高として用いる。5 つの音の音高を選び出す際、抽出したパワースペクトルの局所ピーク群からエネルギーの大きい順に、ハーモニー特徴の 5 つのスロットに格納する。もし 1 つの楽曲片に含まれる局所ピーク数が 5 未満の場合、ハーモニー特徴の残ったスロットは 0 で埋める。例としてある楽曲片が、エネルギーが大きい順にノートナンバー 70、60、80 の 3 つの音を含む場合、ハーモニー特徴は [17, 7, 27, 0, 0] となる。サブ指紋のハーモニー特徴は、5 ビットの音高情報を 5 つ保有するため、25 ビットのバイナリデータで表される。

STFT によって得られたすべてのパワースペクトルからサブ指紋を生成したのち、そのサブ指紋すべてを時系列順に連結して 1 つのオーディオ指紋を生成する。

3.6 オーディオ指紋の比較

楽曲の類似度を計算するため、提案手法では楽曲から生成したオーディオ指紋を比較し、楽曲間距離を算出する。

楽曲間距離の算出は、サブ指紋ブロック単位で行う。図 7 に、サブ指紋ブロックの概要を示す。サブ指紋ブロックとは図 7 に示すように、256 個のサブ指紋の時系列順の連結である。

図 8 に、サブ指紋ブロックの比較の概要を示す。サブ指紋ブロックの比較では、楽曲 A のあるサブ指紋ブロックと、楽曲 B のすべてのサブ指紋ブロックとを比較する。これは、楽曲 A の一部分と類似した部分は、楽曲 B のどこにでも出現する可能性があるからである。始めに、オーディオ指紋をサブ指紋ブロックに分割する。次に、2 つのサブ指紋ブロック間のリズム、メロディ、ハーモニーの特徴をそれぞれ比較する。最後に、3 つの比較の結果を合計し、サブ指紋ブロック間の距離を導出する。

上記の距離導出ステップは、2 つのオーディオ指紋内の

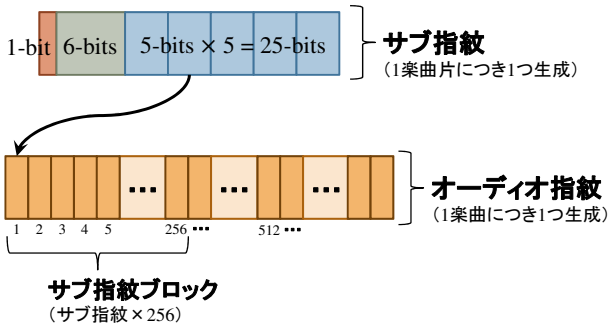

図 7 サブ指紋ブロックの概要.

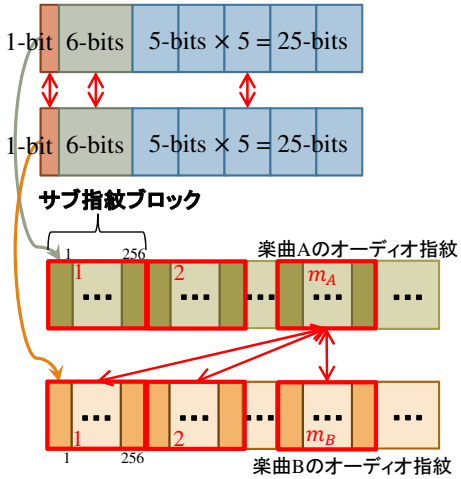

図 8 サブ指紋ブロックの比較.

すべてのサブ指紋ブロックの組み合わせで行う。導出したすべてのサブ指紋ブロック間距離の中で最も小さい距離を、楽曲間距離として用いる。

3.6.1 リズム特徴ビットの比較

提案手法では、サブ指紋ブロック間のリズム距離としてハミング距離を使用する。楽曲 A、B のサブ指紋ブロック間のリズム距離を式 (6) に示す。

$$\sum_{t=1}^{256} \begin{cases} 1, & (rhythm_A(m_A,t) \oplus rhythm_B(m_B,t)), \\ 0, & (otherwise), \end{cases} \tag{6}$$

ここで $rhythm_A(m_A,t)$ は、楽曲 A の m_A 番目のサブ指紋ブロック内の、t 番目のサブ指紋のリズム特徴ビットである。2 つの楽曲の音の発生時刻が完全に同じ場合、それぞれのサブ指紋ブロックのリズム特徴ビットも完全に同じ値となるため、サブ指紋ブロック間のハミング距離は 0 になる。サブ指紋ブロックに含まれるリズム特徴ビットは 256 ビットなので、サブ指紋ブロック間のリズム距離は最大で 256 となる。

3.6.2 メロディ特徴ビットの比較

楽曲 A、B のサブ指紋ブロック間のメロディ距離を式 (7) に示す。

$$\sum_{t=1}^{256} |melody_A(m_A,t) - melody_B(m_B,t)|, \tag{7}$$

ここで $melody_A(m_A,t)$ は、楽曲 A の m_A 番目のサブ指紋ブ

ロック内の、t 番目のサブ指紋のメロディ特徴ビットを整数値に復号したメロディ特徴値である。

楽曲 A、B のサブ指紋ブロック間の、メロディ距離の算出手順を説明する。始めに、それぞれのサブ指紋のメロディ特徴ビットを整数値のメロディ特徴値に復号する。次に、整数値にした各メロディ特徴値の差の絶対値を求める。サブ指紋ブロックは 256 個のサブ指紋の連結なので、この計算を 256 回行うことになる。最後に、上記ステップで計算した差の絶対値を合計し、メロディ距離を導出する。

式 (8) に、メロディ特徴値のとりうる範囲を示す。

$$-30 \leq melody_A(m_A, t) \leq 30. \tag{8}$$

式 (5)、および式 (8) より、ノートナンバーの範囲が 54 から 84 であり、整数値にしたメロディ特徴値は-30 から 30 までの範囲であるため、サブ指紋当たりのメロディ距離は最大で 60 となり、サブ指紋ブロック間のメロディ距離は最大で 15,360 となる。リズム距離との正規化のため、算出したメロディ距離を定数 64 で除す。この定数 64 は、リズム距離の最大値 256 に近くなるよう定めたものである。

3.6.3 ハーモニー特徴ビットの比較

楽曲 A、B のサブ指紋ブロック間のハーモニー距離を式 (9) に示す。

$$\sum_{t=1}^{256} \begin{cases} 0, & (cn(m_A, m_B, t) \geq 3) \\ 3 - cn(m_A, m_B, t), & (otherwise), \end{cases} \tag{9}$$

ここで $cn(m_A, m_B, t)$ は、楽曲 A、B のそれぞれ m_A、m_B 番目のサブ指紋ブロックの、t 番目のサブ指紋のハーモニー特徴における、共通の音高を持つ音の数を示す。コードは少なくとも 3 つの音で構成されていることから、各サブ指紋が共通の音高を持つ音を 3 つ以上有する場合、ハーモニー距離は 0 とする。

サブ指紋当たりのハーモニー距離は最大で 3 なので、サブ指紋ブロック間のハーモニー距離は最大で 768 となる。リズム距離、メロディ距離との正規化のため、算出したハーモニー距離を定数 6 で除す。この定数 6 は、楽曲間距離導出の予備実験により導出された、メロディ距離の平均値に近くなるよう定めたものである。

4. 評価実験と考察

本章では、提案手法の妥当性を検証するため、提案したオーディオ指紋手法を用いて楽曲間の距離を導出する。実験に用いる楽曲は、以下の 52 曲のクラシック音楽からなる。なお、実験に使用する楽曲は、RWC 研究用音楽データベース [22, 23] のクラシック音楽データベース（50 曲）の楽曲を使用した。

A グループ：48 曲

48 曲ともすべて異なるクラシック音楽であり、同じ楽曲は存在しない。楽曲 ID は 1 から 48 である。

表 1　実験に使用するパラメータ.

楽曲データのサンプリング周波数	8,000 Hz
STFT の窓幅	8,192 (1.024 seconds)
STFT のシフト幅	256 (0.032 seconds)
STFT に用いる窓関数	ハミング窓

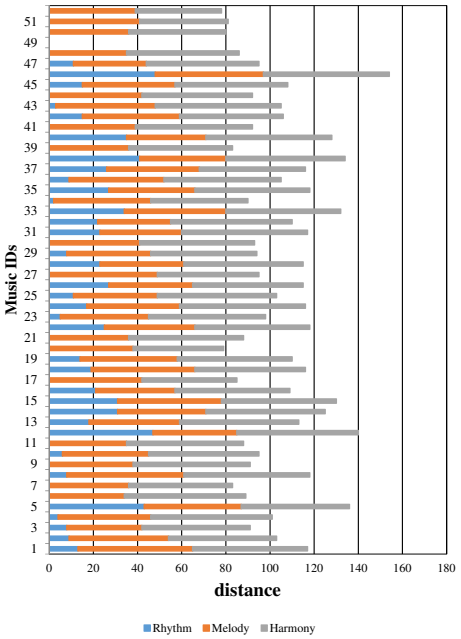

図 9　導出された楽曲間距離.

B グループ：4 曲

4 曲ともすべて同一の楽曲であるが、演奏楽団および録音環境等が異なる。楽曲 ID は 49 から 52 である。

実験に使用したパラメータを表 1 に示す。

評価実験は以下の手順で行う。始めに、実験に使用するすべての楽曲からオーディオ指紋を生成する。続いて、グループ B の楽曲 4 曲の中から 1 曲を無作為に選択する。最後に、選択した楽曲とその他すべての楽曲との楽曲間距離をオーディオ指紋を用いて導出し、その距離を比較する。

4.1 実験結果と考察

図 9 に、実験により導出された楽曲間距離を示す。楽曲間距離はリズム距離、メロディ距離、ハーモニー距離の合計であり、色分けは各音楽的特徴距離の内訳を示す。図 9 の縦軸は使用した楽曲 ID、横軸は楽曲間距離を示す。今回の実験では、参照楽曲として楽曲 ID 49 が選択されている。

図 9 より、楽曲 ID 49 同士の距離が 0 になっている。これは、同一の音響信号からは完全に同じオーディオ指紋が生成されるからである。

参照楽曲同士の比較結果を除くと、リズム距離は 0 から 48 の範囲、メロディ距離は 33 から 53 の範囲、ハーモニー距離は 39 から 57 であり、最大値と最小値の差はそれぞれ 48、20、18 であることから、リズム、メロディ、ハーモニーの中で、最も楽曲間距離に影響しているのはリズムで

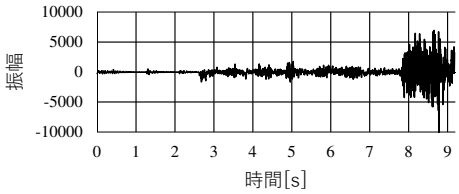

図 10 楽曲 ID 51 の信号（参照楽曲との楽曲間距離最小箇所）.

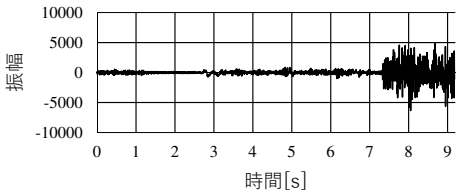

図 11 参照楽曲の信号（楽曲 ID 51 との楽曲間距離最小箇所）.

あることがわかる。また、参照楽曲とのリズム距離が 0 となった楽曲が 52 曲中 17 曲存在した。これは、今回実験に用いた楽曲がすべてクラシックで同一ジャンルに属しており、類似したリズムを持つ楽曲が多く存在したことが原因と考えられる。

図 9 より、参照楽曲と、演奏楽団、および録音環境等のみが異なる同一の楽曲である楽曲 ID 50、51、52 との楽曲間距離は、他の楽曲との距離と比較して短くなっている。このことから提案手法は、演奏楽団、および録音環境が異なる同一楽曲を正しく判定することができることがわかる。

しかし、異なる環境で録音された同一楽曲の楽曲間距離は他の楽曲間距離と比較して短いが、大幅に短いわけではない。これは、同一の楽曲の演奏であっても、演奏楽団や録音環境の差異により多少の違いはやはり存在し、楽譜上で同一の演奏位置を探し出せない場合があるからと考えられる。

次に、楽曲間距離が短い楽曲ペア、長い楽曲ペアについて調査する。楽曲間距離の短い、参照楽曲と楽曲 ID 51 の間で、サブ指紋ブロック間の距離が最小となった部分はそれぞれ、参照楽曲の演奏時間 1 分 19 秒からの 9 秒間と、楽曲 ID 51 の演奏時間 1 分 3 秒からの 9 秒間であった。図 10 と図 11 にそれぞれ、当該演奏区間の波形を示す。

図 10、図 11 より、最も類似していると判定された箇所の波形の形状は類似していることがわかる。実際に、図 10、図 11 の 2 つの音響信号を聞き比べたところ、2 つの音響信号は楽譜上で同一の演奏位置であることが分かった。このことから提案手法は、演奏楽団や録音環境が異なっていても、同一楽曲の、楽譜上で同一の演奏位置を正確に探し出すことができることがわかる。

また、楽曲間距離の長い、参照楽曲と楽曲 ID 46 の間で、サブ指紋ブロック間の距離が最小となった部分はそれぞれ、参照楽曲の演奏時間 2 分 5 秒からの 9 秒間と、楽曲 ID 46 の演奏時間 2 分 18 秒からの 9 秒間であった。図 12 と図 13 にそれぞれ、当該演奏区間の波形を示す。

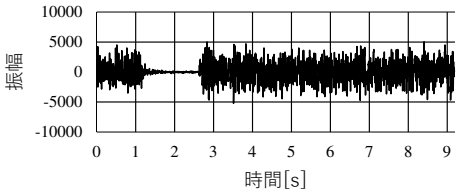

図 12 楽曲 ID 46 の信号（参照楽曲との楽曲間距離最小箇所）.

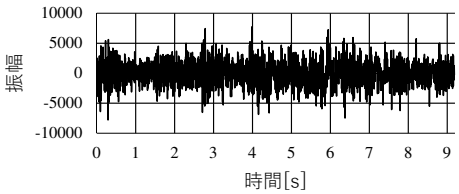

図 13 参照楽曲の信号（楽曲 ID 46 との楽曲間距離最小箇所）.

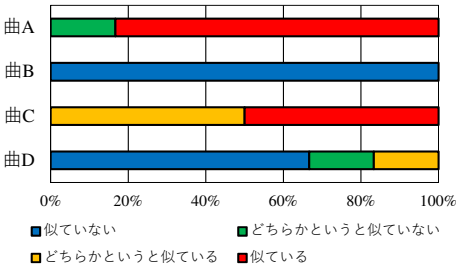

図 14 楽曲 A～楽曲 D の類似度に関するアンケートの結果.

図 10、図 11 より、当該楽曲の波形の形状は大きく異なることがわかる。実際に、図 12、図 13 の 2 つの音響信号を聞き比べたところ、参照楽曲は交響曲であり、楽曲 ID 46 は声楽曲であった。このことから提案手法は、類似していない楽曲同士も正しく判定できることがわかる。

提案手法の導出した楽曲間距離の妥当性を検証するため、第三者による官能テストを行った。提案手法により参照楽曲と似ていると判定された楽曲 2 曲（楽曲 A：ID 51、楽曲 C：ID 50）、および似ていないと判定された楽曲 2 曲（楽曲 B：ID 46、楽曲 D：ID 12）について、サブ指紋ブロック間の距離が最小となった部分の演奏が似ていたかという質問に対し、「似ていない」、「どちらかというと似ていない」、「どちらかというと似ている」、「似ている」の 4 段階で回答するアンケートを行い、6 名から回答を得た。図 14 に、楽曲 A～楽曲 D の類似度に関するアンケートの結果を示す。

楽曲 A と楽曲 C は参照楽曲と同じ楽曲であり、楽曲 A については楽譜上の演奏位置も同一であり、どちらも「似ていない」と回答した人はおらず、「似ている」、または「どちらかというと似ている」と回答した人が 80%を上回った。一方、楽曲 B、楽曲 D については「似ていない」、または「どちらかというと似ていない」と回答した人が 80%を上回り、特に楽曲 B は 100%の人が「似ていない」と回答した。この結果から提案手法は、人が似ていると感じる楽曲、似ていないと感じる楽曲を正しく判別することができたといえる。

5. おわりに

ブロードバンドサービスの普及、オンライン楽曲配信サービスの一般化などにより、多くの人がそうしたサービスを利用して気に入った楽曲を楽しんだり、新しい楽曲を探したりできるようになった。しかし、楽曲配信サービスの保有するデータベースが肥大化し、膨大な数の楽曲が保存されるようになると、その中から新たに気に入る楽曲を人手で探すのは大変な作業となる。そのため、コンピュータを用いた自動楽曲検索の研究が盛んに行われている。

ユーザが気に入る楽曲は、ユーザが現在気に入っている楽曲と類似している可能性が高いと考えられる。つまり、ユーザが所有している気に入った楽曲と類似した楽曲を検索することにより、ユーザの求める楽曲が検索できる可能性がある。多くの研究者が、コンピュータを用いた効率的な類似楽曲検索の研究を行っている。

本稿では、音楽の 3 つの基本要素に焦点を当てた、類似楽曲検索のためのオーディオ指紋の生成方法、および比較方法を提案した。従来のオーディオ指紋は音響的な特徴のみを用いていたため、類似楽曲検索には不向きであった。提案手法では、音響的な特徴から、より抽象的な音楽的特徴を取り出して符号化することで、類似楽曲検索向きのオーディオ指紋を生成する。これは、2 つの楽曲が類似した音楽的特徴を含んでいるとき、2 つの楽曲は類似楽曲といえるからである。

提案手法の妥当性の検証として、提案したオーディオ指紋生成法を用いて実際の楽曲からオーディオ指紋を生成し、オーディオ指紋を用いて楽曲間距離を導出した。また、提案手法により導出された楽曲間距離が短い楽曲、長い楽曲が実際に似ているかについてのアンケート評価を行った。その結果、提案手法が似ていると判定した楽曲について、「似ている」、または「どちらかというと似ている」と回答した人が 80%を上回ったことから、提案手法は類似した楽曲を選び出すことができることがわかった。また、演奏楽団や録音環境が異なる同一楽曲同士の比較では、演奏のテンポの差がある場合においても各楽曲の楽譜上の同一位置を探し出し、類似楽曲と判定することができた。

今後の課題としては、クラシック音楽以外のジャンルにおける楽曲に対する楽曲間距離導出の妥当性の検証、提案手法を用いることによる類似楽曲検索の検索精度の向上の検証などが挙げられる。

参考文献

[1] “ISMIR: International Society of Music Information Retrieval”, http://www.ismir.net/.
[2] “YouTube”, https://www.youtube.com/.
[3] “Google Play Music”, https://play.google.com/store/music.
[4] A. Rauber, E. Pampalk and D. Merkl, “Using Psycho-Acoustic Models and Self Organizing Maps to Create a Hierarchical Structuring of Music by Sound Similarity”, *The International Society of Music Information Retrieval (ISMIR)*, pp. 71–80 (2002).
[5] R. B. Dannenberg and N. Hu, “Understanding Search Performance in Query-By-Humming Systems”, *The International Society of Music Information Retrieval (ISMIR)*, pp. 232–237 (2004).
[6] K. Itoyama, M. Goto, K. Komatani, T. Ogata and H. G. Okuno, “Query-by-Example Music Retrieval Approach Based on Musical Genre Shift by Changing Instrument Volume”, *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)* (2009).
[7] 獅々堀正幹, 大西泰代, 柘植覚, 北研二, “Earth Mover’s Distance を用いたハミングによる類似音楽検索手法”, 情報処理学会論文誌, Vol. 48, No. 1, pp. 300–311 (2007).
[8] 大野和久, 鈴木優, 川越恭二, “楽曲全体における特徴量の傾向に基づいた類似検索手法”, 日本データベース学会論文誌, Vol. 7, No. 1, pp. 233–238 (2008).
[9] J. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System”, *The International Society of Music Information Retrieval (ISMIR)*, pp. 107–115 (2002).
[10] 北研二, 肖清梅, “オーディオ指紋検索に適した高速なハミング空間検索”, 研究報告音楽情報科学 *(MUS)*, Vol. 2011, No. 4, pp. 1–6 (2011).
[11] 澁谷崇, 安部素嗣, 西口正之, “擬似正弦波成分を用いた残響・雑音にロバストなオーディオフィンガープリンティング”, 研究報告音楽情報科学 *(MUS)*, No. 13 (2013).
[12] 黒沢隆朝, 楽典, 音楽之友社 (1947).
[13] 水野正敏, 水野式音楽理論解体新書, 株式会社シンコーミュージック・エンターテインメント (2006).
[14] 松下泰雄, フーリエ解析 基礎と応用, 培風館 (2001).
[15] 佐藤幸男, 信号処理入門（図解メカトロニクス入門シリーズ）, オーム社 (1999).
[16] J. W. Cooley and J. W. Tukey, “An Algorithm for the Machine Calculation of Complex Fourier Series”, *Mathematics of Computation*, Vol. 19, No. 90, pp. 297–301 (1965).
[17] 原裕一郎, 井口征士, “フーリエ変換における窓関数の位相特性”, 計測自動制御学会論文集, Vol. 19, No. 7, pp. 551–556 (1983).
[18] 松田稔, 秋山好一, 森和義, “日本の楽曲の基本的特徴 : 音高について”, 日本音響学会誌, Vol. 50, No. 11, pp. 897–905 (1994).
[19] J. O. Smith III and X. Serra, “PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation”, *International Computer Music Conference*, pp. 290–297 (1987).
[20] M. Abe and J. O. Smith III, “AM/FM Rate Estimation and Bias Correction for Time-Varying Sinusoidal Modeling”, No. STAN-M-118 (2004).
[21] 安部素嗣, ジュリアス・スミス, “FFT の 2 次補間に基づく正弦波パラメータ推定法の設計基準 : 擬似定常な正弦波成分の場合”, 電子情報通信学会技術研究報告. SIP, 信号処理, Vol. 104, No. 306, pp. 7–12 (2004).
[22] “RWC 研究用音楽データベース”, https://staff.aist.go.jp/m.goto/RWC-MDB/index-j.html.
[23] G. Masataka, “Development of the RWC Music Database”, Proceedings of the 18th International Congress on Acoustics (ICA 2004), pp. 553–556 (2004).