情報処理学会論文誌 Vol.54 No.4 1350–1361 (Apr. 2013)

楽曲間の類似判断における個人性データの収集とその分析

川渕 将太[1,a)] 宮島 千代美[1] 北岡 教英[1] 武田 一哉[1]

受付日 2012年7月2日, 採録日 2013年1月11日

概要：楽曲間主観的類似度の個人性を明らかにするためのデータ収集を行った．音楽情報処理の研究で広く用いられている RWC 研究用音楽データベースの「ポピュラー音楽」から選択された 200 の楽曲ペアに関して 27 名の被験者が類似度を評価した．また，各々の被験者は楽曲ペアの全体的な類似度とは別に，メロディ，テンポ・リズム，声質，楽器構成についての類似度も回答した．類似度の回答結果の分析から，「似ている／似ていない」の判断境界が個人ごとに大きくばらつくことが示唆された．また，収集されたデータを用いて個人に最適化された楽曲間距離関数（重み付けユークリッド距離）を学習することで個人ごとに主観的な楽曲間類似度を推定する実験を行った．その結果，距離関数の学習によって「声質」に関する類似性推定の精度が向上したことから，重み付けユークリッド距離を用いた個人適応の効果が明らかになった．

キーワード：楽曲間類似度，主観的類似度，データ収集，個人性

Data Collection for Individuality Analysis of Subjective Music Similarity

Shota Kawabuchi[1,a)] Chiyomi Miyajima[1] Norihide Kitaoka[1] Kazuya Takeda[1]

Received: July 2, 2012, Accepted: January 11, 2013

Abstract: We collected data to analyze the individuality of subjective music similarity. 27 subjects evaluated the similarity of 200 pairs of songs chosen from the RWC "Popular Music Database," a database widely used in the field of music information processing. In addition to overall similarity, each subject also evaluated the similarity of the melody, tempo/rhythm, vocals and instruments for the pairs. Analysis of the collected data suggests that decision boundaries between similar and dissimilar pairs vary widely between individuals. Using the collected data, we trained an optimized distance function (weighted Euclidean distance) between songs for each subject. As a result, the trained distance function improved the precision of similarity estimation for vocals, demonstrating the effectiveness of individual optimization using weighted Euclidean distance.

Keywords: musical similarity, subjective similarity, data collection, individuality

1. はじめに

近年，インターネットを通じた大規模楽曲データベースへのアクセスや，大容量メディアによる個人の大量の楽曲ファイルの利用が可能になった．そのような大量の楽曲から即座に好みの曲を検索したり [1], [2]，ユーザの嗜好に応じた曲を推薦したりするための研究やアプリケーション開発が進められている [3], [4]．そのようなシステムを研究するうえで，楽曲の類似度の定義の仕方について様々な議論がこれまでになされている．楽曲の音響特徴量を利用する手法は，これまで多くの音楽情報検索システムで用いられてきた．音響的類似度の算出の方法は，楽曲からメル周波数ケプストラム係数（Mel-Frequency Cepstrum Coefficient: MFCC）のような，スペクトルの形状に関する特徴量を短時間フレームごとに算出し，その分布間の距離を算出する手法が一般的である．たとえば Pampalk [5] は楽曲から MFCC を短時間フレームごとに算出し，その

[1] 名古屋大学
Nagoya University, Nagoya, Aichi 464–8603, Japan
[a)] shota.kawabuchi@g.sp.m.is.nagoya-u.ac.jp

分布を正規分布もしくは混合正規分布で表現し，その分布間距離を KL 情報量で算出することで音色の類似度とした．音響特徴量として代表的なものには，音色を表すものとして上で述べた MFCC のほかにスペクトル重心やゼロ交差数などスペクトル形状を表す特徴量が，和音を表すものにクロマベクトル，リズムを表すものに Fluctuation Patterns [6] や Rhythm Histogram [7] などがあげられる．

本稿では，楽曲検索に関して，主観的に類似した楽曲を音響的特徴に基づき推定する方法の研究を行う．そのためには，人間の音楽知覚において重要な音響的特徴を抽出し，人間が行うのと同じやり方でそれらの特徴量を比較することが望まれる．しかし，人間の音楽の知覚には明らかでない部分が多く，主観的な類似度を音響的特徴に基づき正確に推定することは非常に困難である．そこで，楽曲間の類似度を人間が主観的に評価したデータを収集し，そのデータを参照することで有効な特徴量および特徴量間類似度を明らかにすることを目指す．収集したデータを用いれば，従来使われてきた音響特徴量・類似度から主観的類似度の推定に最適なものを選択することは可能である．しかし，従来広く用いられてきた単純な音響特徴量を用いる手法には「ガラスの天井」[8] と呼ばれる性能限界があるといわれている．

音響特徴量を用いる手法の性能を向上させるために有効だと考えられる方法の 1 つは，主観的類似度の個人性に着目することである．様々な論文で指摘されているように [4], [9]，主観的類似度を構成する空間は人により異なっていると考えられ，検索システムではユーザに最適な類似尺度を考える必要があると考えられる．たとえば，Novello ら [9] は 9 ジャンル，18 曲の楽曲を用い楽曲類似性の主観評価データを収集した．実験では，被験者に 3 曲の楽曲を提示し，最も似ているペアと最も似ていないペアを選ばせ，102 組の楽曲のペアに対し評価が行われた．収集したデータから楽曲のペアについて類似度の順位を被験者ごとに算出し，被験者間の順位の相関をケンドールの一致係数（Kendall's coefficient of concordance: KCC）により評価した．その結果，楽曲のペアにより KCC の値は大きく異なっていた．つまり，大多数が似ている，もしくは似ていないと評価する楽曲のペアもあれば，意見が分かれるペアもあるということである．このことから Novello らは被験者によって楽曲の類似度を評価する空間が異なる可能性を指摘している．

客観的な評価尺度と主観評価の個人性を関連付ける方法についても研究が行われている．Hoashi ら [3] は，ユーザに好みのジャンルや楽曲を選んでもらい，その情報からユーザの嗜好を表すベクトルを作成することでユーザの好みに合わせ楽曲を推薦する手法を提案している．Vignoli ら [4] は，楽曲の類似度を音色，ジャンル，テンポ，発売年，雰囲気の 5 つの特徴量の重み付け和で表現し，この重みパラメータをユーザが手動で設定することで個人の嗜好をシステムに反映することを試みた．Lampropoulos ら [10] は，音響特徴量を入力としたニューラルネットワークにより類似楽曲を検索するとともに，出力された楽曲群に対してユーザが順位と類似度を与えて再度学習させることでニューラルネットワークを最適化するシステムを提案している．またこの研究では，音楽知覚に影響を与える特徴量は個人により異なるという仮説を立て，特徴量セットからいくつかのサブセットを構築し用いることで最適な特徴量セットをユーザが選べるような仕組みにしている．

しかし，被験者ごとに最適な類似度関数を学習することで楽曲検索システムの性能を向上させる試みは多く報告されているが，主観的類似度において，ユーザ間で具体的にどのような個人差があるかを，共通の楽曲を用いて明らかにする研究はあまり行われていない．さらに，従来多く報告されている，既存の楽曲を用いた研究では，アーティストの情報など，音楽的性質以外の要素が類似性の判断に影響を与えているおそれもある．そこで本研究では，音楽情報検索研究用に広く用いられた楽曲データベースである RWC 研究用音楽データベース「ポピュラー音楽」[11] を使用する．RWC 研究用音楽データベースは研究者が研究目的に利用するうえで，共通利用，学術利用の自由が確保された音楽情報処理研究用のデータベースである．データベースに含まれる楽曲はデータベース作成のために作られた楽曲であり，一般に知られている楽曲や有名なアーティストの楽曲ではないことから，本研究の目的に適している．RWC 研究用音楽データベース「ポピュラー音楽」の楽曲 80 曲の組合せの中から 200 ペアを選び出し，200 ペアに対する楽曲類似性の主観評価データを収集する．さらに収集した主観評価データに基づき，被験者ごとに最適な楽曲間類似度関数を学習する実験を行うことで，楽曲間主観的類似度の個人性が，音響特徴空間上でどのように表現されうるかを明らかにする．

2. 主観的類似度データ収集実験

2.1 被験者

実験に参加した被験者は男性 13 名，女性 14 名の合計 27 名であった．被験者の年代は 20 代であった．被験者 27 名のうち 10 名が楽器経験者であった．

27 名という人数は，楽曲間類似度の参照データとしては必ずしも十分な量ではないが，本実験では被験者の年代を 20 代に限定している．被験者の年代が異なれば類似性判断の傾向も異なり，収集するデータもより多く必要になるものと考えられるが，年代を 20 代に限定したことによって後述するとおり，統計的に有意な分析結果を得ることができた．また，20 代の被験者の実験結果を用いて有用な分析手法を発見できればその手法を他の年代に適用することは可能であると考えている．

2.2 使用楽曲

実験に使用した楽曲は RWC 研究用音楽データベースの「ポピュラー音楽」[11] の楽曲 100 曲のうち，日本のポピュラー音楽（J-Pop）スタイルによる楽曲 80 曲（楽曲番号：No. 1–80）を用いた．楽曲の再生区間は楽曲の一番最初のサビ区間の開始時点から 30 秒間とした．サビ区間の切り出しには RWC 研究用音楽データベースへのアノテーションである AIST アノテーション [12] を用いた．

実験では，被験者に対し使用楽曲 80 曲から 2 曲を選び提示した．以下，この 2 曲を合わせて「楽曲ペア」と呼ぶ．実験において各被験者は 200 対の楽曲ペアを聴き，その各々について類似性の評価を行った．個人性を分析する目的から，全被験者が同じ 200 ペアを評価した．以下では，200 ペアの選択方法について説明する．

2.3 楽曲ペアの選択方法

80 曲の楽曲に対し，楽曲ペアは ${}_{80}C_2$ ペア（=3160 ペア）存在するが実験において 3160 ペアを被験者に聴かせるのは被験者負担が大きく現実的ではない．そこで，実験では 3160 ペアのうちから 200 ペアを選び被験者に提示した．主観的に類似したペア・類似していないペアがどのような音響的類似性を持つかを調べたいため，選択される 200 ペアには様々な組合せが含まれることが望ましい．そこで本研究では，複数の音響的類似度を算出し，算出した類似度を用いてできるだけ様々な特徴のペアが選ばれるよう 200 ペアを選択した．使用する音響的類似度には音色の類似度とリズムの類似度の 2 種類を用いた．楽曲の楽器構成などは音色特徴で，楽曲のテンポやリズムパターンはおおよそリズム特徴で，それぞれ表現できると考えこの 2 つの特徴に基づき楽曲ペアを予備選択した．また，これらの 2 種類の特徴には音楽情報処理分野で広く認知された特徴量（MFCC，FP [6]）があることも，この 2 種類を用いた理由である．音響的類似度を用いた楽曲ペア選択の手続きを以下に示す．まず，3160 ペアに対し音色・リズムの 2 種類の音響的類似度を求め，それぞれの類似度分布のパーセンタイル値に基づき 10×10 の領域に分割した（**図 1**）．各々の領域には，おおよそ同数のペアが含まれる．各グリッドから 2 ペアずつランダムに楽曲ペアを選択することにより，実験に使用する 200 楽曲ペアを得た．ただしこのとき，各楽曲が 200 ペア中でほぼ同じ回数だけ出現するように選択した．

ペア選択のための音色・リズムの音響的類似度は以下により計算した．

音色類似度（MFCC 分布間の KL 情報量）

楽曲間の音色の類似度として，各楽曲から短時間フレームごとの MFCC を抽出し，MFCC の VQ ヒストグラム間の KL 情報量を求めた．MFCC の抽出には MIR toolbox 1.3.2 [13] を使用した．MFCC 抽出の条件を**表 1** に示す．

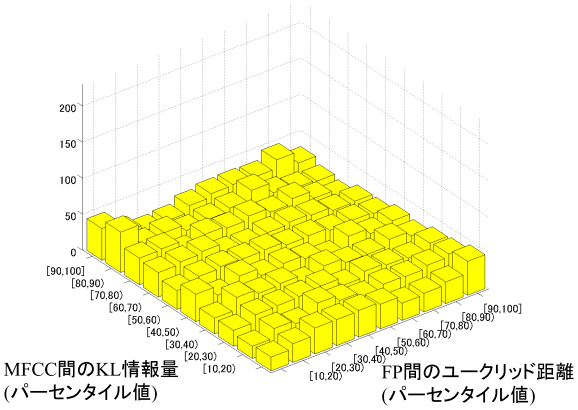

図 1 音響的類似度のパーセンタイル値に関する楽曲ペアの結合ヒストグラム

Fig. 1 Joint histogram of percentile values of acoustical similarities.

表 1 MFCC 抽出の条件

Table 1 Conditions of MFCC extraction.

標本化周波数	44100 Hz
窓関数	Hanning 窓
窓幅	50 ms
シフト幅	25 ms

特徴量としては MFCC の 1–16 次の係数を用いた．また，各時間フレーム n についての時間微分を局所区間の回帰係数

$$\Delta\mathbf{x}(n) = \frac{\sum_{l=-L}^{L} l \cdot \mathbf{x}(n-l)}{\sum_{l=-L}^{L} l^2}$$

として求め特徴ベクトルの成分として加えた．よって，32 次元の短時間特徴量が得られた．回帰係数算出のための計算区間は 150 ms（$L = 2$ に対応）とした．抽出した MFCC をコードブックサイズ 2048 でベクトル量子化し，各楽曲について VQ ヒストグラムを求めた．VQ ヒストグラム間の対称化した KL 情報量

$$D_{SKL}(\mathbf{p}, \mathbf{q}) = D_{KL}(\mathbf{p}, \mathbf{q}) + D_{KL}(\mathbf{q}, \mathbf{p}) \quad (1)$$

を算出し楽曲間の音色の類似度とした．ここで，$\mathbf{p}$ と $\mathbf{q}$ は各楽曲に対し得られた VQ ヒストグラムを表しており，

$$D_{KL}(\mathbf{p}, \mathbf{q}) = \sum_i p(i) \log \frac{p(i)}{q(i)} \quad (2)$$

である．i はヒストグラムのビン番号である．

リズム類似度（FP 間のユークリッド距離）

80 曲からリズム特徴量 Fluctuation Patterns（FP）[6] を抽出し，FP 間のユークリッド距離を算出することで楽曲間のリズムの類似度とした．FP の抽出のためには，まず各時間フレームについて臨界帯域ごとのラウドネス（sone）を算出する（これを Sonogram と呼ぶ）．次に，Sonogram を数秒ごとのセグメントに分割し，時間方向に FFT することにより各臨界帯域についての変調周波数を得る．得られ

た変調周波数の時間方向への差分をとることで各セグメントについてのリズムパターンとする．全セグメントについて同様のリズムパターンを得た後，全セグメントにわたる中央値を要素ごとにとることにより楽曲の FP とする．特徴量の抽出には MA toolbox [14] を用いた．Sonogram 抽出の条件を**表 2** に示す．FP 抽出の条件を**表 3** に示す．特徴抽出の条件については文献 [6] を参考にした．抽出した FP は 20×60 の行列として得られるが，この行列を 1200 次元のベクトルと見なし，そのベクトル間のユークリッド距離を求めることで楽曲間のリズムの類似度とした．

2.4 実験手続き

データ収集実験用に web ブラウザ上で動作する主観評価データ収集システムを構築した．**図 2** にそのインタフェースを示す．被験者に対しシステムは「参照曲」と「評価曲」の 2 曲を提示する．提示する楽曲ペアは 2.3 節で述べた手続きにより選択された 200 ペアの中からランダムな順番で選ばれた．被験者は各楽曲ペアについて評価曲が参照曲と似ているかどうかを 2 段階で評価した．評価を 2 段階としたのは実験の容易性を高めるためである．本研究での大きな目的の 1 つは人手で評価された主観評価データを大量に集めることであり，そのためには実験の手続きをなるべく簡潔にして被験者負担を軽減する必要がある．予備実験において「よく似ている」,「やや似ている」,「まったく似ていない」の 3 段階（カテゴリ形式），および，0.0–10.0 の 100 段階（スコア形式）で評価を行い，選択されたカテゴリとスコアの関係を調べたところ，「よく似ている」ペアに対しては一貫して高いスコアが与えられたのに対し，「やや似ている」と「まったく似ていない」に対しては，与えられたスコアが大きくばらつき，カテゴリの重複もみられた．これらの理由から本研究では 2 段階評価を採用した[*1]．全体的な印象に基づき似ているかどうかを評価した後，楽曲の構成要素（メロディ，テンポ・リズム，声質，楽器構成）ごとに評価を行い，似ていると感じた項目にチェックを入れた．たとえば，被験者が提示された楽曲ペアを似ていないと感じたが，メロディやテンポ・リズムは似ていると感じたのであれば，「似ていない」を選択するとともに，似ている要素欄の「メロディ」と「テンポ・リズム」にチェックを入れた．提示された楽曲ペアを似ていると感じて，テンポ・リズムや楽器構成が似ていたと感じたのであれば，「似ている」を選択し，「テンポ・リズム」と「楽器構成」にチェックを入れた．なお，楽曲は繰り返して再生することができるようにし，各ペアの再評価は被験者の判断で自由に行えるようにした．楽曲の提示にはヘッドホンを用い，楽曲の提示音圧は被験者が好みに応じて変更できるようにした．実験の実施にあたっては 50 分ごとに 10 分間の休憩時間を設けた．

表 2 Sonogram 抽出の条件
Table 2 Conditions of Sonogram extraction.

標本化周波数	11025 Hz
窓関数	Hanning 窓
窓幅	23 ms
シフト幅	12 ms

表 3 FP 抽出の条件
Table 3 Conditions of FP extraction.

窓関数	方形窓
窓幅	6 ms
シフト幅	3 ms

図 2 主観評価データ入力用インタフェース
Fig. 2 Website interface used for data collection.

2.5 収集されたデータ

上記の実験により得られたデータを**図 3**，**図 4** に示す．図 3 は，実験に参加した各被験者が 200 ペア中，似ていると評価したペアの数のヒストグラムである．200 回中 40 回程度似ていると評価する被験者が多い（平均 43.4 回）．要素では，メロディ：平均 38.7 回，テンポ・リズム：平均 61.1 回，声質：平均 56.6 回，楽器構成：平均 41.5 回であった．平均的な被験者が多く存在する一方で，平均から

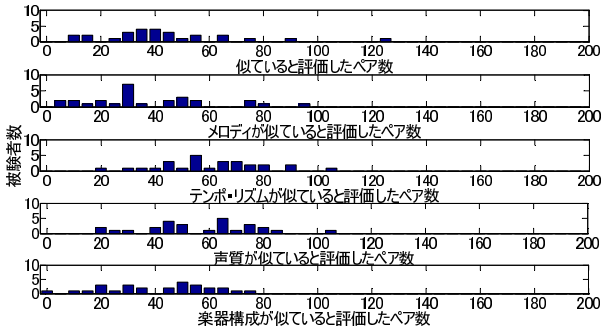

図 3 被験者が似ていると評価したペア数のヒストグラム
Fig. 3 Histograms of number of pairs a subject evaluated as similar.

[*1] これらのことから，被験者は実験においてペアの 2 曲が「似ているか否か」を判断したと考えている．

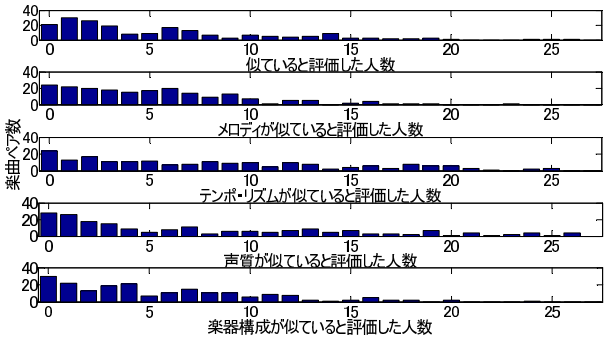

図 4 楽曲ペアを似ていると評価した人数のヒストグラム
Fig. 4 Histograms of number of subjects who evaluated a pair as similar.

大きく外れた値を示す被験者も多数存在することから，似ていると評価する頻度に被験者の個人性が現れることが分かる．すなわち，同程度の類似度を感じても，それを「似ている」と回答するかどうかの判断は個人に依存する．

図 4 は，それぞれの楽曲ペアを似ていると評価した人数のヒストグラムである．評価項目間に若干の傾向の違いはあるが，共通の傾向として少数の被験者にのみ似ていると評価される楽曲ペアが多く，一方で多数の被験者により似ていると評価される楽曲ペアは少数である．

3. 楽曲間音響的類似度の算出

2 章で収集した主観評価データを用いて，楽曲間の主観的類似度によく対応する音響的類似度の算出手法の検討を行う．本研究で検討した，音響類似度算出手法は以下のとおりである．まず，各楽曲の音響信号から短時間特徴量を抽出する．抽出した短時間特徴量を要約することにより楽曲の特徴を表す大局的な特徴ベクトルを得る．2 つの特徴ベクトル間のユークリッド距離を求めることにより楽曲間の類似度を算出する．

以下では，抽出する短時間特徴量とその抽出の条件および，大局的な特徴ベクトルを算出する手法について検討する．

3.1 短時間特徴量の抽出

まず，各々の曲について短時間特徴量を抽出する．抽出された短時間特徴量はメル周波数ケプストラム係数（1–13 次），インテンシティ [15]，スペクトルセントロイド，スペクトルフラックス，スペクトルロールオフ，高周波数エネルギー（brightness と呼ばれることもある）[16] であった．このうち，メル周波数ケプストラム係数，スペクトルセントロイド，スペクトルフラックス，スペクトルロールオフは，ジャンル識別などに有効とされている特徴量を，先行研究 [17] を参考に選出した．また，上記 4 種類の音色特徴量とは別に楽曲のムードを推定する研究で用いられ，有効とされている特徴量としてインテンシティ [15] を選出した．また，予備実験において，MIR toolbox 1.3.2 [13] に含まれる短時間音響特徴量を網羅的に検討した．実験の方法としては，短時間特徴量をベクトル量子化し，各楽曲を短時間特徴量の VQ ヒストグラムとして表現した後，VQ ヒストグラム間の対称化した KL 情報量（式 (1)）を用いて楽曲間の類似度を算出した．算出した楽曲間の類似度と楽曲ペアを似ていると評価した人数との相関をとることで短時間特徴量の性能を評価した．結果，上で述べた 5 種類の特徴量と同等以上の高い相関を示したため，高周波数エネルギー [16] を選出した．

インテンシティとスペクトルフラックス以外の短時間特徴量の抽出には MIR toolbox 1.3.2 [13] を用いた．また，抽出した特徴量の動的変動成分として局所区間における回帰係数を算出し短時間特徴量に加えた．

インテンシティ

インテンシティは楽曲の音量に関する特徴量として提案された [15]．n 番目のフレームにおける振幅スペクトルの総和 $I(n)$ と，$I(n)$ に対するサブバンドの音量の比 $D_j(n)$ からなる特徴量である．

$$I(n) = \sum_{k=0}^{K} X(n,k) \tag{3}$$

$$D_j(n) = \frac{1}{I(n)} \sum_{k=L_j}^{H_j} X(n,k) \tag{4}$$

ここで，$X(n,k)$ は振幅スペクトル，n は時間フレーム番号，k は周波数ビン番号を表す．K はナイキスト周波数に相当する周波数ビン番号を表す．j はサブバンドの番号を表し，サブバンド j の下限の周波数ビン番号を L_j，上限の周波数ビン番号を H_j とする．サブバンドへの振幅スペクトルの分割はオクターブスケールフィルタバンクを用いる．

$$\left[0, \frac{f_s}{2^a}\right), \left[\frac{f_s}{2^a}, \frac{f_s}{2^{a-1}}\right), \ldots, \left[\frac{f_s}{2^2}, \frac{f_s}{2^1}\right] \tag{5}$$

ここで f_s は標本化周波数である．a はサブバンドの数で，本稿では $a = 7$ とした．

スペクトル重心

スペクトル重心は各時刻における振幅スペクトルの重心であり，以下の式により算出する．

$$f_c(n) = \frac{\sum_{k=0}^{K} X(n,k) \cdot k}{\sum_{k=0}^{K} X(n,k)} \tag{6}$$

スペクトルロールオフ

本稿で算出したスペクトルロールオフは，低周波数側から振幅スペクトルの累積和をとった際に，累積和の値が全帯域の総和の 85%に達する周波数であり，以下の式の k_r に対応する．

$$\sum_{k=0}^{k_r} X(n,k) = 0.85 \times \sum_{k=0}^{K} X(n,k) \tag{7}$$

スペクトルフラックス

スペクトルフラックスはスペクトルの非定常性を示す指標であり，本稿では以下のように算出した．

$$F(n) = \frac{1}{K+1}\sum_{k=0}^{K}\left(X(n,k) - X(n-1,k)\right)^2 \qquad (8)$$

高周波数エネルギー（**brightness**）

高周波数エネルギーは brightness とも呼ばれ，スペクトルの特定周波数よりも高い帯域に全体の何%のエネルギーが含まれるかを示す．本稿では 1500 Hz よりも高い帯域に何%のエネルギーが含まれるかを算出した．

動的変動成分

各々の特徴量について，時間フレーム n における時間微分を局所区間の回帰係数として算出した．回帰係数の算出には

$$\Delta\mathbf{x}(n) = \frac{\sum_{l=-L}^{L} l\cdot\mathbf{x}(n-l)}{\sum_{l=-L}^{L} l^2}$$

を用いた．回帰係数算出のための計算区間は 150 ms とした．同様にして二階微分を一階微分の回帰係数として算出した．これらの微係数も短時間特徴量として使用した．

3.2 短時間特徴量の要約

次に，各楽曲について抽出された短時間特徴量を要約し，大局的な特徴ベクトルを得る手法について説明する．本稿では，短時間特徴量の要約手法として 3 つの手法を検討した．そのうちの 2 つはベクトル量子化（VQ）を用いる手法であり，1 つは長時間統計量を用いる手法である．

3.2.1 ベクトル量子化を用いる手法

短時間特徴量をベクトル量子化し VQ ヒストグラムを得る．得られた VQ ヒストグラムは，各次元がヒストグラムの各ビンに対応する特徴ベクトルと見なすことができる．本研究では，この手続きにより得られた特徴ベクトルと，得られた特徴ベクトルの各次元の対数をとることにより変換された特徴ベクトルを検討した．以下，前者を VQ ヒストグラム，後者を対数 VQ ヒストグラムと呼ぶ．また，本稿ではベクトル量子化のクラスタ数として 256，512，1024 を試した．

3.2.2 長時間信号特徴

短時間特徴量から特徴ベクトルを得るもう 1 つの手法として，文献 [17] の手法を用いる．3.1 節で求めた短時間特徴量 $\mathbf{x}(n)$ より，以下の 2 つの値を算出する．

$$m(n,d) = \frac{1}{K}\sum_{k=1}^{K} x(n-k+1,d) \qquad (9)$$

$$s(n,d) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left\{x(n-k+1,d) - m(n,d)\right\}^2} \qquad (10)$$

ただし，$s(n,d)$，$x(n,d)$，$m(n,d)$ はそれぞれ，ベクトル $\mathbf{s}(n)$，$\mathbf{x}(n)$，$\mathbf{m}(n)$ の第 d 成分を表す．$\mathbf{x}(n)$ の各時間フレームに対して求めた $\mathbf{m}(n)$ と $\mathbf{s}(n)$ の各成分の，全時間にわたる平均，標準偏差をとることにより楽曲の特徴ベクトルとした．本稿では，K の値として 20，40，80 を試した．

表 4 特徴ベクトル算出条件
Table 4 Conditions of feature vector calculation.

標本化周波数	44100，22050，16000 Hz
窓関数	Hanning 窓
窓幅	50 ms
シフト幅	25 ms
VQ クラスタ数	256，512，1024
移動平均の点数（秒）	0.5，1.0，2.0

表 5 実験で用いた特徴量セットとその表記
Table 5 Feature sets used in the experiment.

表記	特徴量セット
INS	インテンシティ
TMB	スペクトル重心，スペクトルロールオフ，スペクトルフラックス，高周波数エネルギー
MFC	MFCC
ALL	3.1 節で求めた全特徴量
x+Δ	一階微分を加えたもの
x+Δ+ΔΔ	一階微分と二階微分を加えたもの
PCA	ALL+Δ+ΔΔ を主成分分析したもの

3.3 実験 1（特徴ベクトルの検討）

3.2 節で求めた特徴ベクトルの性能を比較する実験を行った．ここで，特徴ベクトルの性能とは，特徴ベクトル間のユークリッド距離が，楽曲間の主観的類似度と対応する度合いを意味する．

本稿では，各楽曲ペアについて似ていると評価した被験者数および，メロディ/テンポ・リズム/声質/楽器構成が似ていると評価した人数を数え，特徴ベクトル間のユークリッド距離のマイナスをとった値との相関をとることにより特徴ベクトルの性能を評価することとする．実験で比較した特徴ベクトルの算出条件を**表 4** に，特徴量セットとその表記を**表 5** にそれぞれ示す．

実験の結果を**表 6**，**表 7**，**表 8** に示す．表 6，7 はそれぞれ 3.2.1 項で述べた VQ ヒストグラム，対数 VQ ヒストグラムについて，表 5 の特徴量セットの中で，類似していると答えた人数との相関が最も高かった特徴量上位 3 つまでを，構成要素ごとに示している．また，表中の「特徴量セット」は表 5 で示した特徴量セットのどれを用いたか，「f_s」は特徴抽出に用いた音楽データの標本化周波数（Hz），「クラスタ数」はベクトル量子化のクラスタ数，「相関」は似ていると評価した人数との相関をそれぞれ示している．

表 6 似ていると評価した人数との相関が高い条件（VQ ヒストグラム）

Table 6 Conditions of feature vectors which achieved a high correlation to the number of subjects who evaluated a pair as similar (VQ histogram).

評価の観点	特徴量セット	f_s	クラスタ数	相関
全体	TMB+Δ+ΔΔ	16000	256	0.33
	TMB+Δ+ΔΔ	22050	256	0.30
	TMB	16000	256	0.30
メロディ	TMB+Δ+ΔΔ	16000	256	0.30
	TMB+Δ+ΔΔ	22050	256	0.28
	TMB+Δ	22050	256	0.25
テンポ・リズム	TMB+Δ+ΔΔ	16000	256	0.30
	TMB+Δ+ΔΔ	22050	256	0.27
	TMB+Δ+ΔΔ	16000	512	0.27
声質	TMB	16000	256	0.16
	TMB	16000	512	0.16
	TMB	22050	256	0.15
楽器構成	TMB+Δ+ΔΔ	16000	256	0.40
	TMB+Δ+ΔΔ	22050	256	0.36
	TMB+Δ+ΔΔ	16000	512	0.36

表 7 似ていると評価した人数との相関が高い条件（対数 VQ ヒストグラム）

Table 7 Conditions of feature vectors which achieved a high correlation to the number of subjects who evaluated a pair as similar (logarithmic VQ histogram).

評価の観点	特徴量セット	f_s	クラスタ数	相関
全体	PCA	16000	512	0.57
	PCA	44100	512	0.57
	PCA	22050	512	0.56
メロディ	INS+Δ+ΔΔ	44100	512	0.49
	INS+Δ+ΔΔ	44100	256	0.49
	TMB+Δ+ΔΔ	16000	256	0.47
テンポ・リズム	INS+Δ+ΔΔ	44100	256	0.44
	INS+Δ+ΔΔ	22050	512	0.44
	INS+Δ+ΔΔ	44100	512	0.43
声質	MFC+Δ	44100	256	0.43
	MFC+Δ	44100	512	0.43
	MFC+Δ	44100	1024	0.42
楽器構成	ALL+Δ+ΔΔ	22050	256	0.68
	ALL+Δ+ΔΔ	44100	256	0.68
	PCA	44100	256	0.67

表中の「全体」以下の 3 項目は似ていると評価した被験者数との相関が最も高かった特徴ベクトルを，「メロディ」以下の 3 項目はメロディが似ていると評価した被験者数との相関が最も高かった特徴ベクトルをそれぞれ表している．以下，「テンポ・リズム」「声質」「楽器構成」についても同様である．表 8 では 3.2.2 項で述べた長時間統計量について特徴量と算出の条件を比較している．表中の「特徴量セット」，「f_s」，「相関」は表 6，7 と同じ内容を示しており，「K（秒）」は式 (9) と式 (10) における K の値に相当する時間を示している．「類似性」「メロディ」「テンポ・リズム」「声質」「楽器構成」についても表 6，7 と同様である．実験の結果より，構成要素ごとに最適な音響特徴量やその算出条件，特徴ベクトルの構成手法が異なるということが分かる．

表 8 似ていると評価した人数との相関が高い条件（長時間統計量）

Table 8 Conditions of feature vectors which achieved a high correlation to the number of subjects who evaluated a pair as similar (long-term statistics).

評価の観点	特徴量セット	f_s	K（秒）	相関
全体	ALL+Δ+ΔΔ	16000	0.5	0.44
	TMB+Δ+ΔΔ	16000	0.5	0.44
	ALL+Δ	16000	0.5	0.44
メロディ	ALL+Δ+ΔΔ	16000	0.5	0.43
	TMB+Δ+ΔΔ	16000	0.5	0.43
	ALL+Δ+ΔΔ	16000	2.0	0.43
テンポ・リズム	ALL+Δ	16000	2.0	0.39
	ALL+Δ+ΔΔ	16000	2.0	0.39
	TMB+Δ+ΔΔ	16000	2.0	0.39
声質	MFC+Δ+ΔΔ	16000	0.5	0.15
	MFC+Δ+ΔΔ	22050	0.5	0.15
	MFC+Δ	16000	0.5	0.15
楽器構成	MFC+Δ+ΔΔ	16000	0.5	0.57
	MFC+Δ+ΔΔ	22050	0.5	0.57
	MFC+Δ+ΔΔ	44100	0.5	0.57

VQ ヒストグラム（表 6）の場合は，全評価項目に対し，使用する特徴量として音色特徴量（スペクトル重心，スペクトルロールオフ，スペクトルフラックス，高周波数エネルギー）を用い，声質以外の項目に対しては一階微分と二階微分を特徴量に加えたときに相関が高い．相関については，楽器構成に対し 0.40 と相対的に高い相関が得られた．それ以外の評価項目については類似性，メロディ，テンポ・リズムで 0.30 程度，声質では 0.16 と相関が低い．

対数 VQ ヒストグラム（表 7）の場合は，類似性と楽器構成において全特徴量に一階微分と二階微分を加えた特徴量，もしくはそれを主成分分析した特徴量，メロディとテンポ・リズムではインテンシティに一階微分と二階微分を加えた特徴量，声質では MFCC に一階微分を加えた特徴量で相関が高い．相関については類似性と楽器構成においてそれぞれ 0.57，0.68 と相対的に高い値が得られ，その他の項目においても 0.40 以上と，VQ ヒストグラムよりも全体的に高い値が得られている．特に，声質において VQ ヒストグラムとの違いが顕著である．

長時間統計量（表 8）の場合は，類似性，メロディ，テンポ・リズムにおいて，全特徴量もしくは音色特徴量に一階微分・二階微分を加えた特徴量で高い相関が得られる傾向にあり，声質と楽器構成では MFCC に一階微分と二階微分を加えた特徴量において相対的に高い相関が得られている．K の値は 0.5 秒のときに高い相関が得られる傾向にあ

るが，テンポ・リズムについては 2.0 秒の場合に高い相関が得られている．相関については，VQ ヒストグラム，対数 VQ ヒストグラム同様，楽器構成について 0.57 と比較的高い相関が得られるが，VQ ヒストグラム同様，声質については 0.15 と相関が低く，類似性，メロディ，テンポ・リズムについては 0.40 前後の値が得られている[*2]．

以上の結果から，対数 VQ ヒストグラムが他の長時間特徴量に比べて相対的に高い相関を示しており，この 3 種類の中では主観的類似度の推定に最も適していると考えることができる．対数 VQ ヒストグラムでは，対数をとることで，少頻度でしか出現しない VQ プロトタイプ群の出現頻度の偏りが，相対的に強調される．したがって，対数 VQ ヒストグラムは，楽曲全体にわたって流れる平均的な旋律や音色だけでなく，短時間でも特徴的な旋律や音色を類似性の計算に考慮することができる．対数 VQ ヒストグラムの優位性はこの性質によるものと考えているが，その詳細な検証は今後の課題である．また，短時間特徴の要約方法が変わると，高い相関値を示す特徴量の組合せが変化する理由も，平均的な楽曲区間に関連した音響特徴量と，短時間でも特徴的な楽曲区間に関連した音響特徴とが異なることが原因と考えられる．

4. 距離関数の個人適応

3 章では，主観評価データを用いて楽曲間の主観的類似度によく対応する音響的類似度の算出手法を検討した．しかし，主観的類似度と音響的類似度の対応関係は個人ごとに異なると考えられるため，個人ごとに最適化された音響的類似度を求めることができると望ましい．そこで，本研究では以下の方法により音響的類似度の個人適応を行う．3 章で算出した特徴ベクトル間の距離を以下に示す重み付けユークリッド距離により算出する．

$$||\mathbf{v}_i - \mathbf{v}_j||_{\mathbf{W}} = \sqrt{(\mathbf{v}_i - \mathbf{v}_j)^T \mathbf{W} (\mathbf{v}_i - \mathbf{v}_j)} \qquad (11)$$

ただし，$\mathbf{v}_i$ と $\mathbf{v}_j$ は D 次元特徴ベクトルであり，$\mathbf{W}$ は $D \times D$ の半正定値行列である．重み行列 $\mathbf{W}$ を各々の被験者について最適化することにより，被験者ごとに異なる距離関数を得ることができる．本稿では，主観評価データを用いて重み行列を最適化することにより被験者ごとに最適な距離関数を得た．

以下では，重み行列の最適化手法について述べた後，実験を行い個人適応の効果を確認する．

[*2] 対数 VQ ヒストグラムと「似ている」と評価した人数との相関値は評価観点ごとに差があるものの，0.4–0.7 程度であり，両者に線形関係があると結論するのは困難である．本章の目的は，次章以降で距離計算を利用者ごとに適応させる方法を検討する前段階として，音響分析条件や 3 種類の短時間特徴量要約手法（VQ ヒストグラム，対数 VQ ヒストグラム，長時間統計量）の性能を，27 名全員のデータを用いて比較することであり，対数 VQ ヒストグラムに基づく尺度で，「似ている」と評価する人数が予測できると結論づけるものではない．

4.1 距離学習

2.5 節の結果から，同一の類似度に感じられる楽曲ペアであっても，それを「似ている」と判断する境界は被験者によって異なると考えられる．このため，主観評価データは被験者が似ていると判断する境界の違いとどのような音響的特徴を重視するかという 2 種類の個人性を反映していると考えられる．

しかし，個人性を学習するうえでより重要なのはどのような音響的特徴を重視するかであると考えられる．重視する音響的特徴の個人差が大きいのであれば，ユーザごとにまったく異なる楽曲を提示する必要があるからである．したがって，本稿では，どのような音響的特徴を重視するかのみを反映させた距離関数を学習することを目指す．具体的には，「似ている」とラベル付けされたペアに対しては値が小さくなり，「似ていない」とラベル付けされたペアに対しては値が大きくなるような重み付けユークリッド距離（式 (11)）を学習することにより距離関数の最適化を実現する[*3]．

最適な距離関数を学習する手法としては，距離学習（metric learning）と呼ばれる機械学習の手法を用いることで実現できる．本研究では，そのうち，「ペアが似ているか否か」のデータに対し適用可能な手法として以下にあげる MLR と ITML を選択して使用した．

4.1.1 Metric learning to rank

Metric Learning to Rank（MLR）[18] は Structural SVM [19] に基づく距離学習手法である．この手法は最適な距離関数を順序尺度の観点から学習する．つまり，学習された距離関数に基づき楽曲を並べ替えると，ある楽曲に対し「似ている」とラベル付けられた楽曲が「似ていない」とラベル付けられた楽曲よりも序列の上位に位置付けられるような距離関数を学習する．

そのような距離関数を実現するために，以下の制約付き最小化問題を解く．

$$\begin{aligned} \min_{\mathbf{W} \succeq 0, \xi} \quad & \mathrm{tr}(\mathbf{W}) + C \cdot \xi \\ \text{s. t.} \quad & \forall i: \ \langle \mathbf{W}, \mathbf{\Psi}(\mathbf{x}_i, \mathbf{y}_i) \rangle_F \\ & \qquad \geq \langle \mathbf{W}, \mathbf{\Psi}(\mathbf{x}_i, \mathbf{y}) \rangle_F + \Delta(\mathbf{y}_i, \mathbf{y}) - \xi \end{aligned} \qquad (12)$$

ただし，ξ はスラック変数であり，$\mathrm{tr}(\mathbf{A})$ は行列 $\mathbf{A}$ のトレース，C はスラック変数のトレードオフを調整するパラメータ，$<>_F$ はフロベニウス内積を表し，$\langle \mathbf{A}, \mathbf{B} \rangle_F = \mathrm{tr}(\mathbf{A}^T \mathbf{B})$ である．$\mathbf{x}_i$ は i 番目の特徴ベクトル（本稿では，各々の楽曲が 1 つのベクトルとして表現され，総楽曲数が 80 曲であるから，$i = 1, ..., 80$），$\mathbf{y}$ は $\mathbf{x}_i$ に対する（$\mathbf{x}_i$ を除く）各曲の類似度の序列，特に，$\mathbf{y}_i$ は $\mathbf{x}_i$ についての正解の序

[*3] 主観評価実験において，被験者が感じる類似度を数値として直接計測できれば音響特徴量との関係を直接分析可能であるが，そのような主観量を直接安定して大量に計測することは実験技術上必ずしも容易ではない．

列（「似ている」ベクトルが「似ていない」ベクトルよりも順位が高い）を表している．行列 $\mathbf{\Psi}(\mathbf{x}_i, \mathbf{y})$ は特徴マップ（feature map）と呼ばれる，これはベクトル $\mathbf{x}_i$ と序列 $\mathbf{y}$ との関係を表す．$\Delta(\mathbf{y}_i, \mathbf{y})$ は序列間について定義される損失関数であり，典型的には $[0, 1]$ の値をとる．

望ましい距離関数を学習するために，特徴マップ $\mathbf{\Psi}(\mathbf{x}_i, \mathbf{y})$ は以下のように定義される．

$$\mathbf{\Psi}(\mathbf{x}_i, \mathbf{y}) = \sum_{\mathbf{x}_S} \sum_{\mathbf{x}_D} y_{sd} \left(\frac{\mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_D) - \mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_S)}{|N_S| \cdot |N_D|} \right) \tag{13}$$

ただし，$\mathbf{x}_S$ は $\mathbf{x}_i$ に「似ている」ベクトル，$\mathbf{x}_D$ は「似ていない」ベクトルをそれぞれ表し，$|N_S|$ と $|N_D|$ は「似ている」ベクトルの数と，「似ていない」ベクトルの数をそれぞれ表す．また，

$$y_{sd} = \begin{cases} +1 & \mathbf{x}_S \text{ is ranked higher than } \mathbf{x}_\mathrm{D} \text{ in } \mathbf{y} \\ -1 & \mathbf{x}_S \text{ is ranked lower than } \mathbf{x}_\mathrm{D} \text{ in } \mathbf{y} \end{cases} \tag{14}$$

であり，

$$\mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_j) = -(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{15}$$

である．以上のように定義された特徴マップを用いることにより，望ましい重み行列 $\mathbf{W}$ は $\mathbf{y}$ が誤った序列のときよりも正しい序列 $\mathbf{y}_i$ のときに高い得点 $\langle \mathbf{W}, \mathbf{\Psi}(\mathbf{x}_i, \mathbf{y}_i) \rangle_F$ をとる．

損失関数 $\Delta(\mathbf{y}_i, \mathbf{y})$ は以下のように定義される．

$$\Delta(\mathbf{y}_i, \mathbf{y}) = \mathrm{Score}(\mathbf{y}_i) - \mathrm{Score}(\mathbf{y}) = 1 - \mathrm{Score}(\mathbf{y}) \tag{16}$$

ただし，$\mathrm{Score}(\mathbf{y})$ は $[0, 1]$ の値をとる関数であり，序列 $\mathbf{y}$ が正しい序列に近いほど高い値をとる（序列が完璧な場合 1 をとる）．したがって，損失関数は序列の誤りの指標である．本稿では，$\mathrm{Score}(\mathbf{y})$ として後述する ROC 曲線の下の面積（Area Under the ROC Curve; AUC）を用いた．AUC は「似ている」と回答した回答数には依存しない指標であるため，本研究が目指すモデル構築に適した基準である．

MLR によって最適な距離関数を学習するために，MATLAB により実装された MLR（mlr-1.0）[*4] を用いた．

4.1.2 Information theoretic metric learning

Information Theoretic Metric Learning（ITML）[20] は，「似ている」ペアに対しては距離関数の値がある上限より小さく，「似ていない」ペアに対しては距離関数の値がある下限よりも小さくなるような制約を満たすもののうち，何らかの行列 $\mathbf{W}_0$ にできるだけ近くなるように重み行列 $\mathbf{W}$ 学習する手法である．本稿では単位行列 $\mathbf{I}$ にできるだけ近くなるように重み行列 $\mathbf{W}$ を学習する．つまり学習の結果得られる重み付けユークリッド距離 $||\mathbf{x}_i - \mathbf{x}_j||_\mathbf{W}$ が重み付けを行わないユークリッド距離にできるだけ近くなるよう重み行列を学習することに相当する．前項同様，この距離学習においても，「似ている」と判定した回数は距離学習に直接影響を与えない．

そのような距離関数を実現するために，以下の制約付き最小化問題を解く，

$$\begin{aligned} \min_{\mathbf{W} \succeq 0, \boldsymbol{\xi}} \quad & D_{ld}(\mathbf{W}, \mathbf{I}) + C \cdot D_{ld}(diag(\boldsymbol{\xi}), diag(\boldsymbol{\xi}_0)) \\ \text{s. t.} \quad & ||\mathbf{x}_i - \mathbf{x}_j||_\mathbf{W} \leq \xi(i,j) \quad (i,j) \in S \\ & ||\mathbf{x}_i - \mathbf{x}_j||_\mathbf{W} \geq \xi(i,j) \quad (i,j) \in D \end{aligned} \tag{17}$$

ただし，$\mathbf{I}$ は $\mathbf{W}$ と同じ大きさの単位行列であり，S と D はそれぞれ「似ている」ペアと「似ていない」ペアの集合である（つまり，$(i,j) \in S$ は $\mathbf{x}_i$ と $\mathbf{x}_j$ のペアが「似ている」とラベル付けられていることを意味している）．$\boldsymbol{\xi}$ はスラック変数のベクトルであり $\boldsymbol{\xi}_0$（各要素が各制約に対応し，「似ている」ペアに対しては距離関数の上限値 u，「似ていない」ペアに対しては下限値 l をとる）によって初期化される．本稿では，学習データ内における距離関数値の分布の 5 パーセンタイル値と 95 パーセンタイル値を式 (17) の下限 l，上限 u としてそれぞれ用いた．D_{ld} は LogDet 情報量と呼ばれ，以下の式により定義される．

$$D_{ld}(\mathbf{A}, \mathbf{B}) = \mathrm{tr}(\mathbf{A}\mathbf{B}^{-1}) - \log\det(\mathbf{A}\mathbf{B}^{-1}) - d \tag{18}$$

ただし，$\mathbf{A}$ と $\mathbf{B}$ は $d \times d$ の行列である．

ITML を用いて最適な距離関数を学習するために，MATLAB により実装された ITML（itml-1.2）[*5] を用いた．

4.2 実験 2（距離関数の個人適応）

距離関数の個人適応により主観的類似度の推定性能をどの程度改善できるかを確認するための実験を行った．3.3 節の結果に基づき，特徴ベクトルには対数 VQ ヒストグラムを用いた．特徴ベクトル算出の条件を**表 9** に示す．距離学習手法として 4.1 節で紹介した 2 種類を用いた．

類似ラベルとして 2 章で述べた主観的類似度評価データを用い，各被験者について最適な距離関数を学習した．実験は 10 分割交差確認法により行われた，つまり，80 曲を 8 曲ずつの 10 組に分け，そのうち 9 組（72 曲）を学習に

表 9 特徴ベクトル算出の条件
Table 9 Conditions of feature vector calculation.

標本化周波数	16000 Hz
窓関数	Hanning 窓
窓幅	50 ms
シフト幅	25 ms
特徴量セット	PCA（表 5 参照）
クラスタ数	512

[*4] http://cseweb.ucsd.edu/~bmcfee/code/mlr/

[*5] http://www.cs.utexas.edu/~pjain/itml/

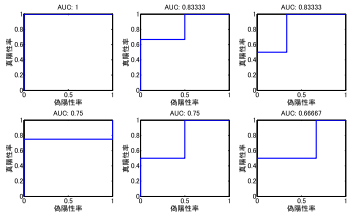

図 5 ROC 曲線とそれに対応する AUC の例
Fig. 5 Examples of ROC curves and corresponding AUC values.

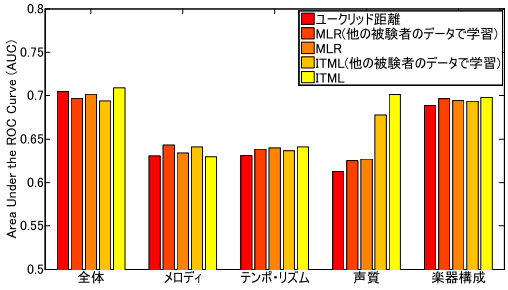

図 6 各評価観点・条件における AUC の値
Fig. 6 AUC values for each component and condition.

用い，残りの 1 組（8 曲）を評価に用いた．この手続きを評価用の組を変えながら 10 回繰り返し，10 回の平均により全体の性能を評価した．

学習された距離関数の性能評価には ROC 曲線の下の面積（Area Under the ROC Curve; AUC）を用いた．具体的な計算方法としては，ある楽曲・ある被験者について縦軸が真陽性率，横軸が偽陽性率の ROC 曲線を算出しその下の面積を求める．全楽曲・全被験者について同様に AUC を算出しその平均をとることにより距離関数を評価した．ROC 曲線と，それに対応する AUC の例を**図 5** に示す．図 5 はある 6 曲に対し，ある被験者の評価データを用いて ROC 曲線を描いたものである．各グラフの上に示されているのは，その ROC 曲線に対して求めた AUC の値である．

距離学習により主観的類似度推定性能がどの程度改善されたかを確認するために，重み付けを行わないユークリッド距離を比較対象として用いた．距離関数の個人性を確認する目的で，学習した距離関数を学習に用いた被験者以外のデータによっても評価した．

MLR もしくは ITML を用いて距離学習を行ううえで，スラック変数のトレードオフパラメータを $C \in \{10^{-2}, 10^{-1}, \ldots, 10^{6}\}$ と変化させた．結果では，各学習アルゴリズムの各項目について最も高い AUC の値を示した C を用いた．最適な距離関数の学習には，総合的な類似度の評価データだけでなく，各構成要素（メロディ，テンポ・リズム，声質，楽器構成）についての評価データも用いた．

4.3 実験結果

実験 2 の結果を**図 6** に示す．図 6 は評価観点ごとに 5 つの条件（ユークリッド距離，他の被験者のデータを使って MLR により学習された距離，同一被験者のデータを使って MLR により学習された距離，他の被験者のデータを使って ITML により学習された距離，同一被験者のデータを使って ITML により学習された距離）で AUC を算出し，その全被験者にわたる平均をとったものを示している．声質以外の評価観点においては，学習により得られた距離関数（MLR，ITML）と重み付けを行わないユークリッド距離とに違いがないことが分かる．この結果から，声質以外の観点では，主観的に感じる類似度と音響特徴量との対応に個人間で大きな相違がないことが考えられる．よって，声質以外の観点では「似ている／似ていない」の判断境界が個人性においてより重要である．声質については，重み付けを行わないユークリッド距離に比べ，学習により得られた距離関数を用いた場合で高い AUC が得られている．また，他の被験者のデータで評価した場合に比べ，適応した被験者で評価した場合に高い AUC が得られたことから，声質の類似度に対応する物理量は，人によって一様でなく，個人ごとの距離学習が効果的であったと考えられる．声質について，ユークリッド距離を用いた群と ITML を用いて本人に適応した群の AUC の値について対応のある t 検定（片側検定）を行ったところ，2 群間に 1%の有意差が認められた．ITML を用いて他者に適応した群と本人に適応した群の AUC の値について同様に検定を行ったところ，こちらも 2 群間に 1%の有意差が認められた．これにより，声質に関する類似度においては個人ごとの距離学習が効果的であったことが示された．なお，声質以外の観点に関しては，いずれも個人適応前後の性能に 1%の有意差は認められなかった．

声質の類似度は，アーティストの選択・推薦に重要であり，声質の類似度と対応の良い物理尺度が被験者ごとに学習可能なことは，実用上重要な結果である．しかし．重み行列の学習により個人性を学習することができたことについて．合理的な説明をするためには．各要素の主観的類似度を規定する物理的特徴を示し．声質に関しては複数の物理量が関係していることを示す必要があると考えられる．本稿ではこのことについて取り上げないが．これは今後の重要な課題の 1 つである．

5. 結論

本稿では，まず楽曲間主観的類似度の参照データとなるデータ収集実験について報告した．収集されたデータを分析することにより，主観評価に個人性が存在することが確認された．特に，被験者の個人性は似ていると評価する頻度において顕著であった．

次に，収集したデータを用いて楽曲間の主観的類似度の

推定に適した特徴ベクトルの算出手法・条件について検討した．VQ ヒストグラム，対数 VQ ヒストグラム，長時間統計量という 3 種類の特徴ベクトルを検討したが，楽曲ペアを似ていると評価した人数との線形相関により評価したところ，いずれの評価観点に対しても，対数 VQ ヒストグラムを特徴ベクトルとして用いたときに最も良い結果が得られた．特に，楽器構成を観点とした類似度については 0.68 と最も高い相関が得られた．次に，検討により得られた特徴ベクトルを用いて，楽曲間類似度の個人性を表現する方法を検討した．具体的には，重み付けユークリッド距離の重み行列を被験者ごとに学習することによりこれを行った．実験の結果から，声質についてみた場合，個人適応を行った場合とそうでない場合で結果に有意差が認められ，声質に関しては，主観的類似度と物理的特徴との対応が個人によって異なることが示唆された．

ここで得られた結果は，個人ごとに最適化された楽曲検索・推薦に利用可能である．楽曲検索がかかえる問題の 1 つに，いわゆるコールドスタートの問題があり，再生回数が少ない新しい楽曲を適切なユーザに提示するためには，音響信号処理のみを用いて，特定ユーザのプレイリストとの主観的な類似度を評価することなどが必要である．これに関して，本研究では，ボーカルの声質を重視して類似する曲を検索する場合には，個人ごとに異なる重み行列を用いて音響尺度を計算することで，より良い結果が得られることを実験的に示すことができた．

主観的な楽曲類似度の理解において，残された重要な課題は，個人ごとに異なる「似ている/似ていない」の判断境界を考慮したモデルの構築である．本稿で扱った主観的類似度のモデルは，楽曲間の主観的な類似度が，複数の音響尺度とどのように関係しているかを分析する方法を与えるものであり，類似度がどの程度だと似ていると判断されるかということについては考えていない．しかし，2.5 節でみたように，似ていると判断する境界は個人により大きく異なると考えられる．構築したデータベースでは，27 名の被験者が同一の楽曲ペア群に対して「似ている/似ていない」の判断を行っており，今後構築したデータベースを用いて個人ごとの類似判断のモデルを構築する必要がある．また，本稿で用いられた音響特徴量は楽曲のスペクトル形状を表現するもののみであるが，音楽の知覚においてはリズムや和音，コード進行，メロディなど，これらの特徴量では表現が困難なものも重要であると考えられる．したがって，今後はそれらの特徴量をモデルに取り入れることも考えていく．

今回収集した主観評価実験データは，RWC 研究用音楽データベースを利用した種々の研究に共用するため，http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SSimRWC/において研究目的に公開する．

謝辞 本研究の一部は科学技術振興機構 CREST，および科研費 No.23650088 によるものである．

参考文献

[1] Raubel, A., Pampalk, E. and Merkl, D.: Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity, *The 3th International Conference on Music Information Retrieval* (*ISMIR 2002*), pp.71–80 (2002).

[2] Goto, M. and Goto, T.: Musicream: New music playback interface for streaming, sticking, sorting, and recalling musical pieces, *The 6th International Conference on Music Information Retrieval* (*ISMIR 2005*), pp.404–411 (2005).

[3] Hoashi, K., Matsumoto, K. and Inoue, N.: Personalization of user profiles for content-based music retrieval on relevance feedback, *ACM Multimedia*, pp.110–119 (2003).

[4] Vignoli, F. and Pauws, S.: A music retrieval system based on user-driven similarity and its evaluation, *The 6th International Conference on Music Information Retrieval* (*ISMIR 2005*), pp.272–279 (2005).

[5] Pampalk, E.: Computational models of music similarity and their application in music information retrieval, Ph.D. Thesis, Vienna University of Technology (2006).

[6] Pampalk, E., Rauber, A. and Merkel, D.: Content-based organization and visualization of music archives, *ACM Multimedia*, pp.1–6 (2002).

[7] Lidy, T. and Rauber, A.: Evaluation of feature extractions and psycho-acoustics transformations for music genre classification, *The 6th International Conference on Music Information Retrieval* (*ISMIR 2005*), pp.34–41 (2005).

[8] Aucouturier, J. and Pachet, F.: Improving timbre similarity: How high is the sky?, *Journal of Negative Results in Speech and Audio Sciences*, Vol.1, No.1, pp.1–13 (2004).

[9] Novello, A., McKinney, M. and Kphlrausch, A.: Perceptual evaluation of music similarity, *The 7th International Conference on Music Information Retrieval* (*ISMIR 2006*), pp.246–249 (2006).

[10] Lampropoulos, A., Sotiropoulos, D. and Tsihrintzis, G.A.: Individualization of music similarity perception via feature subset selection, *IEEE International Conference on Systems, Man and Cybernetics*, Vol.1, pp.552–556 (2004).

[11] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. 3rd International Conference on Music Information Retrieval* (*ISMIR 2002*), pp.287–288 (2002).

[12] Goto, M.: AIST annotation for the RWC music database, *Proc. 7th International Conference on Music Information Retrieval* (*ISMIR 2006*), pp.359–360 (2006).

[13] Lartillot, O. and Toiviainen, P.: MIR in Matlab (II): A toolbox for musical feature extraction from audio, *Proc. 8th International Conference on Music Information Retrieval* (*ISMIR 2007*), pp.127–130 (2007).

[14] Pampalk, E.: A MATLAB toolbox to compute similarity from audio, *The 5th International Conference on Music Information Retrieval* (*ISMIR 2004*) (2004).

[15] Lu, L., Liu, D. and Zhang, H.J.: Automatic mood detection and tracking of music audio signals, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.14, No.1,

pp.5–18 (2006).
[16] Juslin, P.N.: Cue utilization in communication of emotion in music performance: Relating performance to perception, *Journal of Experimental Psychology: Human Perception and Performance*, Vol.26, No.6, pp.1707–1813 (2000).
[17] Tzanetakis, G.: MARSYAS submissions to MIREX 2010, *The 6th Music Information Retrieval Evaluation eXchange* (*MIREX 2010*) (2010).
[18] McFee, B. and Lanckriet, G.: Metric learning to rank, *Proc. 27th Annual International Conference on Machine Learning*, Fürnkranz, J. and Joachims, T. (Eds.), Haifa, Israel, pp.775–782 (2010).
[19] Tsochantatidis, I., Joachims, T., Hofmann, T. and Aitun, Y.: Large Margin Methods for Structured and Interdependent Output Variables, *Journal of Machine Learning Research*, Vol.6, pp.1453–1484 (2005).
[20] Davis, J.V., Kulis, B., Sra, S. and Dhillon, I.S.: Information-theoretic metric learning, *Proc. International Conference on Machine Learning*, Corvallis, Oregon, USA, pp.209–216 (2007).

川渕 将太

1988 年生．2010 年愛媛大学法文学部人文学科卒業．2012 年名古屋大学大学院情報科学研究科修士課程修了．修士（情報科学）．2012 年同博士課程入学．類似楽曲検索の研究に従事．日本音響学会学生会員．

宮島 千代美

1973 年生．1996 年名古屋工業大学工学部知能情報システム学科卒業．1998 年同大学大学院修士課程修了．2001 年同博士課程修了．博士（工学）．2001 年名古屋工業大学助手．2003 年名古屋大学助手．2007 年同大学助教．主に運転行動信号処理・音声情報処理の研究に従事．IEEE，電子情報通信学会，日本音響学会，自動車技術会各会員．

北岡 教英

1969 年生．1994 年京都大学大学院工学研究科修士課程修了．1994 年（株）デンソー入社．2000 年豊橋技術科学大学大学院工学研究科博士後期課程修了．博士（工学）．2001 年豊橋技術科学大学助手．2003 年同大学講師．2006 年名古屋大学助教授．2007 年同大学准教授．主として音声認識，音声対話，音声インタフェースに関する研究に従事．IEEE，IEEE-SPS，ISCA，電子情報通信学会，日本音響学会，人工知能学会各会員．

武田 一哉

1960 年生．1983 年名古屋大学工学部電気工学科卒業．1985 年同大学大学院修士課程修了．以降，ATR，KDD 研究所，名古屋大学にて音声合成・認識，音響・音楽信号処理，行動信号処理の研究に従事．IEEE，電子情報通信学会，日本音響学会，自動車技術会各会員．