楽曲間の主観的類似度と音響的類似度との関連付けに関する検討*

◎平賀悠介 (名大)，大石康智 (NTT)，原直，武田一哉 (名大)

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所有が可能になった．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上で，楽曲の類似度の定義の仕方についても様々な議論がこれまでになされている．音響的特徴の類似度を楽曲の類似度として用いる手法は多くの音楽情報検索システムで用いられており，代表的な特徴量としてメル周波数ケプストラム係数（Mel-Frequency Cepstrum Coefficient: MFCC），スペクトル形状に基づく特徴量などが挙げられる．

一方で，楽曲の類似性について人間が主観的に評価したデータを用い楽曲の類似度を表現する研究も行われている．Novello ら [5] は一対比較法により収集した楽曲類似性の主観評価データから楽曲群の相対的な距離を算出し，さらに多次元尺度構成法（Mulutidimensional scaling: MDS）により二次元平面上に楽曲群の相対的な位置関係を表現した．

また，楽曲間類似度の “ground-truth” を構築する試みもなされている．国際音楽情報検索会議 ISMIR の音楽情報検索システムのコンテスト（MIREX）では，いくつかのタスクにおいて楽曲間類似度評価システム “Evalutron 6000” [6] を利用し楽曲検索システムの主観的な性能評価を行っている．評価は，システムが “query” の楽曲 1 曲とそれに対する “candidates” の楽曲数曲を提示し，被験者がそれぞれの “candidate” が “query” に似ているかどうかを 3 段階評価（“NOT Similar”, “Somewhat Similar”, “VERY Similar”）と 100 段階評価（0〜10 点，0.1 点刻み）で評価を与えるというものである．この主観評価データを収集し，Jones ら [7] は楽曲間類似度データベースを構築している．使用している楽曲データベースは “Audio Music Similarity and Retrieval” タスクの場合 7000 曲 10 ジャンルと大きいものであるが，楽曲データは商用楽曲データであるため，データベースを研究者は自由に利用することはできない．楽曲の類似性の判断は主観に基づくものであるため，主観評価データを用い構成された楽曲類似尺度は楽曲の類似性を真に表していると言える．しかし，主観的な評価を与えた楽曲群しか類似度が算出出来ない，といった問題点がある．

最後に，客観的な評価尺度と楽曲の類似性に対する主観評価や個人の嗜好を関連付ける研究を紹介する．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 と Pauws[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 [8] 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．また彼らは，個人の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．

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 [9, 10] の楽曲 202 曲の全ての組み合わせに対する，楽曲類似性の主観評価データを収集し，楽曲間類似度の “ground-truth” を構築する．さらに収集した主観評価データに基づき，楽曲の類似性に対する人の主観的な判断を推定できるような楽曲類似尺度を構築することを目指す．

本稿ではまず，楽曲間類似度主観評価実験の概要を述べる．次に，楽曲の主観的な類似度は音響的な類似度の重み付け和により表されると仮定し，その重み係数を収集した主観評価データに基づき線形回帰モデルにより推定する実験について述べる．

2 主観評価データ収集実験

我々は楽曲間類似度の “ground-truth” を与えるデータベースを構築するための主観評価実験を行っている．web ブラウザ上で動作する主観評価データ収集システム（Fig. 1）を構築し，web を通して実験を行っている．楽曲を再生し，主観評価データを入力するためのインタフェースは Flash 形式のアプリケーションで，FlashPlayer がインストールされている PC であれば使用環境を問わず実験を行うことができる．

2.1 使用楽曲

被験者に提示する楽曲は，RWC 研究用音楽データベースの「ポピュラー音楽」[9] の楽曲 100 曲と「音楽ジャンル」[10] の楽曲 102 曲で構成された計 202 曲を用い，同一曲同士を除く全ての楽曲ペアについて実験を行う．楽曲の再生区間は先行研究 [7] と同様に，曲の中間の前後 15 秒ずつを切り出した 30 秒の区間を用いる．

---

*A study of relationship between a subjective similarity and acoustic similarities of music. by HIRAGA, Yusuke(Nagoya University), OHISHI, Yasunori(NTT), HARA, Sunao, TAKEDA, Kazuya (Nagoya University)

Table 3　被験者の構成（() 内は楽器経験者）

年齢	～19	20～29	30～39	40～49	50～59	60～
男性	3 (0)	30 (6)	47 (11)	16 (4)	16 (3)	17 (1)
女性	2 (0)	43 (15)	61 (22)	26 (5)	13 (0)	6 (0)

Table 1　評価基準の項目

メロディ（C1）	歌やその曲で主となる楽器のメロディライン
テンポ（C2）	曲の速さ
雰囲気（C3）	曲の印象．繊細，力強い，など
ジャンル（C4）	音楽のジャンル
声質（C5）	ボーカルの声質
楽器構成（C6）	使用されている楽器
その他（C7）	上記以外の要素

Table 2　収集実験の条件

楽曲データベース	RWC 研究用音楽データベース： ポピュラー音楽 100 曲 音楽ジャンル 102 曲
総楽曲ペア数	40602 ペア
評価楽曲ペア数	200 曲/人
楽曲再生区間	中間の前後 15 秒ずつ
ファイル形式	MP3, 192kbps, stereo

Fig. 1　主観評価データ入力用インタフェース

2.2　実験手順

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（以後，参照曲と評価曲のペアを「楽曲ペア」と呼ぶ）．各ペアが同一回数評価されるように設計した．被験者は各楽曲ペアについて評価曲が参照曲に似ているか似ていないかと，どのような観点で評価したか（以後「評価基準」と呼ぶ）をシステムに入力する．評価基準は複数回答を可能として，該当する項目を全て選択するように指示している．例えば被験者は提示された楽曲ペアのメロディと雰囲気が異なるために似ていないと感じたのであれば,「似ていない」を選択し,「メロディ」と「雰囲気」にチェックを入れる．評価基準として提示した項目を Table 1 に示す．なお，楽曲は繰り返して再生することができるようにし，実験の休憩，再評価，再ログインは被験者の判断で自由に行えるようにした．また，上記のデータの他に被験者の年齢，性別，職業，日常的な楽器演奏経験の有無についてアンケートを行っている．被験者の年齢，性別，楽器経験者数の構成を Table 3 に示す．

2.3　収集したデータの予備分析

2009 年 7 月 23 日時点における収集データの概要を Table 4 に示す．

Fig. 2 に 8520 ペアの楽曲ペアが似ていると評価された回数の分布を示す．誰にも似ていると評価されなかったペアが最も多く，以降似ていると評価された回数が増加するに従って指数関数的に楽曲ペア数は減少している．

Table 4　収集したデータの概要

被験者数	271 名
評価ペア数（のべ）	52,609 ペア
評価ペア数（異なり）	8,520 ペア
一人当たり平均評価ペア数	194.1 ペア

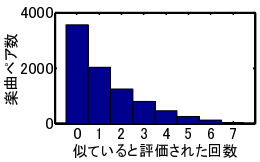

Fig. 2　似ていると評価される回数の分布

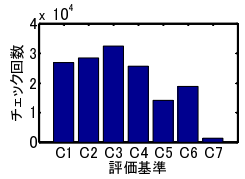

Fig. 3　評価基準の出現頻度

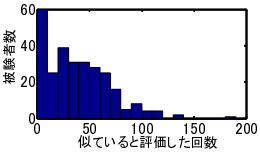

Fig. 4　似ていると評価する回数の分布

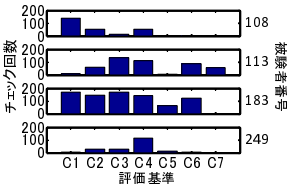

Fig. 5　被験者別に見た評価基準の出現頻度

Fig. 3 は，各評価基準が選択された頻度を示す．ただし評価基準の記号 C1～C7 はそれぞれ「メロディ」,「テンポ」,「雰囲気」,「ジャンル」,「声質」,「楽器構成」,「その他」を表す．最も選ばれた回数が多かったのは「雰囲気」（C3）で，逆に「その他」（C7）を除いて最も回数が少なかったのは「声質」（C5）であった．これは，使用した楽曲データベース中にはボーカルを含まないインストゥルメンタル曲が多数含まれているためであると考えられる．

Fig. 4 に被験者が似ていると評価した回数の分布と，Fig. 5 に各被験者の選んだ評価基準のヒストグラムの例をそれぞれ示す．似ていると感じる頻度や基準は，被験者により大きく異なっていることが確認できる．

3　主観評価と音響的類似度の関連付け

本節では収集した楽曲類似性の主観評価データと楽曲の音響的類似度の関連付け手法について述べる．

3.1　主観評価スコア

本研究では主観的な楽曲間類似度を，主観評価スコア y として次のように定義する．

$$y_i = \frac{\text{楽曲ペア } i \text{ を似ていると評価した被験者数}}{\text{楽曲ペア } i \text{ を評価した被験者数}} \quad (1)$$

y_i は楽曲ペア i の主観評価スコアである．例えば，5 人中 3 人が似ていると判定した楽曲ペア i の主観評価スコア y_i は，$y_i = 3/5 = 0.6$ である．

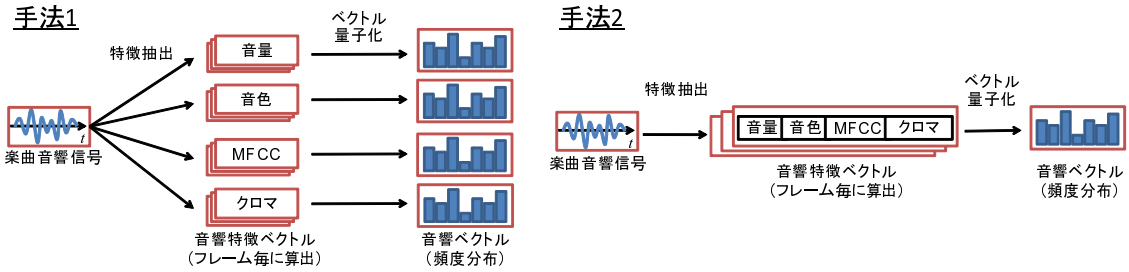

Fig. 6　音響ベクトルの作成

3.2　音響的類似度

本研究では楽曲の音響的特徴を表す音響ベクトルを作成し，そのベクトル間類似度を音響的類似度として定義する．以下では音響特徴量の抽出と音響ベクトルの作成および類似度の算出について述べる．

3.2.1　音響特徴ベクトルの抽出

楽曲を短時間フーリエ変換（STFT）して求めた短時間パワースペクトルから以下の音響特徴量を算出し，音響特徴ベクトルを作成する．STFT の分析条件を Table 5 に示す．音響特徴量は，音量（インテンシティ [11]），音色 [11]，MFCC（16 次），和音を表す特徴量であるクロマ特徴量（HPCP [12]）の 4 種類を用いる．音量，音色，MFCC については動的変動成分として 1 次の回帰係数

$$\Delta \mathbf{x}(t) = \frac{\sum_{k=-L}^{L} k \cdot \mathbf{x}(t-k)}{\sum_{k=-L}^{L} k^2} \tag{2}$$

を求め，特徴ベクトルの成分として加えた．$\mathbf{x}$ は音響特徴量，L はフレーム幅を表し，本研究では $L = 2$ とした．音量，音色，MFCC，クロマ特徴量の動的変動成分を含めた次元数はそれぞれ 16，50，32，12 である．なお，インテンシティについては STFT を行う前に楽曲信号をサンプリング周波数 44.1kHz から 16kHz にダウンサンプリングした．

本研究ではフレームごとに音響特徴ベクトルを作成する．なお，音響特徴ベクトルは，楽曲データセット全体から算出した平均と分散を用いて正規化した．

3.2.2　音響ベクトルの作成と類似度算出

次に，3.2.1 節で求めた音響特徴ベクトルをベクトル量子化し，頻度分布を音響ベクトルとする．ベクトル量子化のためのコードブックは，全楽曲から求めた音響特徴ベクトルの集合から LBG アルゴリズムを用いクラスタリングし，セントロイドを代表ベクトルとして作成した．コードブックサイズは 2048 とした．こうして求めた音響ベクトルのコサイン類似度を楽曲間の音響的類似度として用いる．

なお，本研究では Fig. 6 に示すように，手法 1，手法 2 の 2 種類の音響ベクトル作成法を検討した．手法 1 では音響特徴ごとに音響ベクトルを作成し，各音響特徴についてそれぞれ類似度を算出する．一方手法 2 では全音響特徴を表す音響ベクトルを作成し，類似度を算出する．したがって各楽曲ペアについて，手法 1 では類似度が 4 つ，手法 2 ではただ 1 つ算出される．

Table 5　STFT の分析条件

サンプリング周波数	44.1kHz
フレーム長	50ms
フレームシフト	25ms
窓関数	ハニング窓

3.3　関連付け手法

3.1 節で求めた主観評価スコア y と 3.2 節で求めた音響的類似度を線形回帰モデルで関連付ける．音響特徴 m についての楽曲ペア i の音響的類似度を d_{mi} とすると，主観評価に基づく新しい類似度 $\hat{y}_i$ は

$$\hat{y}_i = b + w_1 d_{1i} + \cdots + w_M d_{Mi} \tag{3}$$

で表される．b はバイアス項，w_m は偏回帰係数で，最小二乗法により学習される．M は音響特徴ベクトル数で，手法 1 では $M = 4$，手法 2 では $M = 1$ となる．手法 1 は各音響的特徴に重みを与えたモデル，手法 2 は重みを与えないモデルとなる．

4　関連付け性能の評価実験

音響特徴量の類似度と主観評価による楽曲の主観的類似度との関連付け性能を検証した．

4.1　実験条件

収集した主観評価データのうち，5 名以上に評価された楽曲ペア $N = 7399$ ペアについての評価値を用いて評価実験を行った．楽曲 7399 ペアで出現した楽曲は 202 曲である．偏回帰係数の値を比較するため，音響的類似度は平均と分散を用いて正規化した．評価は 7398 ペアを偏回帰係数の学習に用い，残り 1 ペアを評価に用いる leave-one-out 法で行った．

実験結果の評価基準には平均二乗誤差（Root Mean Square Error: RMSE）

$$RMSE = \sqrt{\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2} \tag{4}$$

を用いた．従属変数である主観評価スコア y の標準偏差よりも RMSE の値が低ければ，提案したモデルは従属変数の分散を説明できている，すなわちモデルが有効であるといえる．

4.2　実験結果と考察

RMSE の値を Table 6 に示す．主観評価スコアの標準偏差よりも RMSE が小さいことから，モデルの有効性が確認された．また，手法 2 に比べ手法 1 の RMSE は小さくなった．偏回帰係数の値を Table 7 に

Table 6　主観評価スコアの標準偏差と RMSE 値

主観評価スコアの標準偏差	0.2328
手法 1 の推定誤差	**0.1967**
手法 2 の推定誤差	0.2026

Table 7　偏回帰係数の値

バイアス項	0.2198
音量	0.0403
音色	0.0944
MFCC	0.0199
クロマ	0.0423

Table 8　音響的類似度の相関行列

	音量	音色	MFCC	クロマ
音量	1.0000			
音色	**0.7352**	1.0000		
MFCC	0.4475	0.3788	1.0000	
クロマ	0.3004	0.3841	0.1752	1.0000

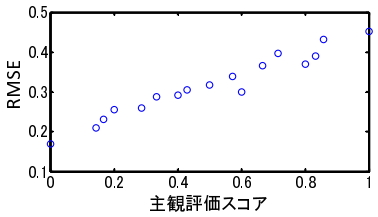

Fig. 7　主観評価スコアごとの推定誤差

示す．音色の特徴量に強く重み付けされており，楽曲の類似度に対する音色の特徴量，即ちスペクトル形状に基づく特徴量の重要性が確認された．Fig. 7 は主観評価スコアごとの推定誤差を示す．主観評価スコアが高くなる程，推定誤差は大きくなった．これは，実験に用いたデータの主観評価スコアに偏りがあるため，主観評価スコアの低いデータにフィッティングされてしまったためであると考えられる．

ところで線形回帰モデルは，各説明変数は互いに独立であることを仮定する．しかし，Table 8 で示すように，各類似度の相関係数 r を調査したところ，各説明変数には相関があることが確認され，特に音量と音色で $r = 0.7937$ と高い相関が認められた．本稿では式（1）に示したように，主観的類似度が音響的類似度の重み付け和で表されると仮定したが，主観的類似度の定義や音響的類似度との関連付け手法についてはさらなる検討が必要である．

5　まとめと今後の展開

楽曲類似性の主観評価データ収集のためのシステムを構築し，収集実験を行った．収集されたデータの性質を分析し，多様なデータが得られることを確認した．また，主観評価データと楽曲の音響的類似度を線形回帰モデルにより関連付ける手法を検討した．類似度算出実験の結果，提案手法の有効性を確認した．

今後は，全ての楽曲ペアに対する主観評価を得るため，主観評価データ収集実験を継続する．また，主観評価と音響的類似度の関連付けについてさらなる検討を行う．さらに個人ごとにモデル化を行い，個人の音楽知覚に適応した楽曲検索システムの構築を目指す．

参考文献

[1] A. Rauber, *et al.*, “Using Psycho-Acoustic Models and Self-Organizing Maps to Create a Hierarchical Structure of Music by Sound Similarity,” *Proc. ISMIR*, pp.71–80, 2002.

[2] M. Goto and T. Goto, “Musicream: New Music Playback Interface for Streaming, Sticking, and Recalling Musical Pieces,” *Proc. ISMIR*, 2005.

[3] K. Hoashi, *et al.*, “Personalization of User Profiles for Content-based Music Retrieval on Relevance Feedback,” *Proceedings of ACM Multimedia 2003*, pp.110–119, 2003.

[4] F. Vignoli and S. Pauws, “A Music Retrieval System Based on User-Driven Similarity and its Evaluation,” *Proc. ISMIR*, 2005.

[5] A. Novello, *et al.*, “Perceptual evaluation of music similarity,” *Proc. ISMIR2006*, pp.246–249, 2006.

[6] A.A. Gruzd, *et al.*, “Evalutron 6000: Collecting Music Relevance Judgments,” *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007.

[7] M.C. Jones, *et al.*, “Human similarity judgements: Implications for the design of formal evaluations,” *Proc. ISMIR2007*, 2007.

[8] A.S. Lampropoulos, *et al.*, “Individualization of Music Similarity Perception via Feature Subset Selection,” *IEEE, International Conference on Systems, Man and Cybernetics 2004* (2004).

[9] M. Goto, *et al.*, “RWC Music Database: Popular, Classical, and Jazz Music Databases,” *ISMIR*, pp.287–288, 2002.

[10] M. Goto, *et al.*, “RWC Music Database: Music Genre Database and Musical Instrument Sound Database,” *ISMIR*, pp.229–230, 2003.

[11] L. Lu, *et al.*, “Automatic Mood Detection and Tracking of Music Audio Signals,” *IEEE Transactions on audio, speech, and language processing*, vol. 14, No. 1, 2006.

[12] E. Gómez, “Tonal description of music audio signals,” Ph.D. thesis, UPF, Barcelona, Spain, 2006.

[13] O. Lartillot and P. Toiviainen, “MIR in Matlab (II): A toolbox for musical feature extraction from audio,” *Proc. ISMIR*, 2007.