

歌詞と音響特徴量を用いた 楽曲印象軌跡推定法の設計と評価^{*1}

西 川 直 毅^{†1} 糸 山 克 寿^{†1} 藤 原 弘 将^{†2,†3}
後 藤 真 孝^{†2} 尾 形 哲 也^{†1} 奥 乃 博^{†1}

本稿では、歌詞と音響信号のそれぞれが持つ印象の時間変化を印象軌跡として推定し、その二つの組み合わせで楽曲全体の印象軌跡を表現する手法について述べる。歌詞の印象軌跡は、確率的潜在意味解析 (PLSA) を用いて、歌詞中の単語から歌詞の印象を表すトピックを推定することで求める。音響信号の印象軌跡は、重線形回帰分析を用いて音響特徴量から推定する。評価実験では、「The Beatles」の 175 曲の印象軌跡を推定し、それらを複数のクラスにクラスタリングして分析した。各クラスごとの音響特徴量の比較、ソーシャルタグと印象軌跡の比較から、推定された印象軌跡は適切であり、楽曲印象の時間変化が表現できる事がわかった。

Design and Evaluation of a Musical Mood Trajectory Estimation Method Using Lyrics and Acoustic Features

N. NISHIKAWA,^{†1} K. ITOYAMA,^{†1} H. FUJIHARA,^{†2,†3}
M. GOTO,^{†2} T. OGATA^{†1} and H. G. OKUNO^{†1}

This paper describes a method that represents an overall musical mood trajectory (time-varying impressions) of a song by using two mood trajectories estimated for both of the lyrics and audio signals. The mood trajectory of the lyrics is obtained by using the Probabilistic Latent Semantic Analysis (PLSA) to estimate topics (representing impressions) from words in the lyrics. The mood trajectory of the audio signals is estimated from acoustic features by using the Multiple Linear Regression Analysis. In our experiments, mood trajectories of 175 songs by The Beatles are estimated and clustered into several classes. Comparison of acoustic features within each class and comparison of social tags and mood trajectories showed that the estimated mood trajectories is suitable and can represent time-varying impressions of songs.

1. はじめに

近年の楽曲再生ソフトウェアや携帯音楽プレーヤーの音楽ライブラリ容量増加に伴い、音楽情報検索 (MIR: Music Information Retrieval) 技術の需要が高まり、研究がさかんに行われるようになった。例えば、スペクトルなどを表現する音響的特徴の類似度を使用した楽曲検索などがある^{1),2)}。これらの研究は、「人間は音響的特徴が似た楽曲を似ていると知覚する」という観点に基づいていると考えられる。それに対して、人間は音楽を印象の表現として知覚する³⁾との観点からは、楽曲類似度の知覚は音響的特徴ではなく楽曲から受ける印象に基づくと考えられる。つまり、印象を定義し、その印象に基づいて楽曲を分類・検索する機能が実現できれば、MIR はより人間の感性に適した検索結果を提供する事ができると期待される。一方で楽曲の印象推定の従来研究⁴⁾⁻¹⁵⁾はいまだ発展途上である。これは、印象が視聴する個人の主観で評価されるものであり、定量化が困難なためである³⁾。

我々は、楽曲印象を表現する上で以下の 2 点が重要であると考ええる。

- 歌詞と音響信号を両方用いること
- 印象の時間変化を考慮すること

前者については、歌詞は楽曲の内容を言語により表現したもので、楽曲の印象に大きな影響を与えると考えられるためである。実際の楽曲においては、(明るい曲調だが歌詞の内容が暗い場合など) 歌詞と音響信号で受ける印象が異なる場合があり、このような楽曲の印象を適切に表現するには両方を用いることが不可欠である。歌詞と音響信号の両者を用いた従来研究⁹⁾⁻¹¹⁾においても、両者を用いた場合に印象の認識率が向上する例が報告されている。後者については、音楽から受ける印象は楽曲の進行に応じて時々刻々と変化する^{12),13)}と考えられるためである。例えば、(いわゆるサビや A メロなどの) 楽曲の任意の区間によって曲調が大きく異なる楽曲や、歌詞が物語のような時間的構造を持つ楽曲などが存在し、そのような楽曲では楽曲全体で 1 つの印象を推定するだけでは不十分である。

楽曲印象推定の従来研究⁴⁾⁻¹⁵⁾では、これらの 2 点を同時に扱ったものは無かった。例え

^{*1} 本研究は、科研費基盤研究 (S) の支援を受けた。

^{†1} 京都大学 大学院情報学研究科 知能情報学専攻

Dept. of Intelligence Science and Technology, Grad. School of Informatics, Kyoto University

^{†2} 産業技術総合研究所

National Institute of Advanced Industrial Science and Technology (AIST)

^{†3} School of Electronic Engineering and Computer Science, the Queen Mary University of London

ば歌詞音響信号を同時に用いた研究例^{9)–11)}は1つの楽曲について1つの印象を推定する目的で開発されており、楽曲中の印象の時間変化は考慮されていなかった。一方、印象の時間変化を考慮した研究^{12),13)}では歌詞は用いられていなかった。

本稿では時間変化する楽曲印象の時系列、すなわち楽曲印象軌跡の推定手法について述べる。我々は、歌詞と音響信号の2要素が重要との立場から、歌詞と音響信号の印象軌跡をそれぞれ推定し、楽曲全体の印象をその2つの組み合わせとして表現する。このように楽曲印象を多面的に表現することで、上述の歌詞と音響信号で印象が異なる楽曲や複雑な時間構造を持つ楽曲印象などを正確に表現することが可能になる。時間的に変化する印象を計算機上で定量的に扱うため、快-不快を表す Valence 軸と興奮-弛緩を表す Arousal 軸からなる Russell の V-A 平面¹⁶⁾を導入し、印象を V-A 平面上の軌跡として表現する。これにより、歌詞・音響印象推定とは入力楽曲の歌詞と音響信号からそれぞれの印象を表現する V-A 平面上の軌跡を推定する問題となる。

歌詞印象軌跡の推定には、印象表現語を事前知識に用いた制約付き確率的潜在意味解析 (PLSA: Probabilistic Latent Semantic Analysis¹⁷⁾) を用いる。通常の PLSA では、推定される文章 (歌詞) のトピックが必ずしも印象を表現したものにはならないという問題があったが、本研究では事前知識を用いることでトピックを強制的に V-A 平面にマッピングすることが可能にする。音響印象軌跡のためには、先行研究の手法^{12),13)}を参考に、重回帰分析を用いて音響特徴量と印象の関係を学習する。

評価実験として、推定された二つの印象軌跡を用いて楽曲のクラスタリングを行い、各クラスが持つ音響特徴量の比較、各クラスの印象軌跡とソーシャルタグの比較を行った。歌詞と音響信号の類似度合いでのクラスタリング結果と各クラスの印象軌跡とソーシャルタグの比較から、提案法により適切な印象軌跡が推定でき、楽曲印象の時間構造が表現できる事が確認できた。

2. 歌詞と音響特徴量を用いた楽曲印象軌跡推定法

本研究の目的は、楽曲印象の時間的な変化を推定することである。歌詞が既知の音楽音響信号を入力とし、その楽曲の印象の時間変化を表す軌跡を出力する。原理的には歌詞が含まれる限りあらゆるジャンルの音楽に対して適用可能であるが、本稿の評価実験ではポピュラー音楽を使用した。

2.1 印象軌跡の定量化

印象軌跡推定に先立ち、印象軌跡の定量化を行う。まず楽曲の音楽的なまとまり (一般

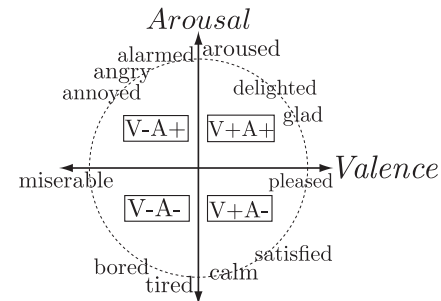


図1 Russell の V-A 平面

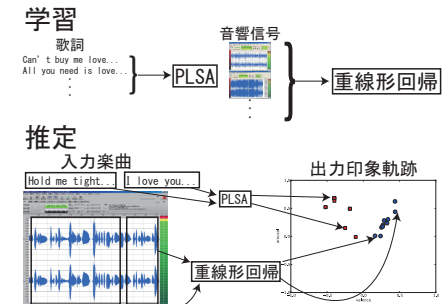


図2 本稿での処理

的にメロ、サビなどと呼ばれる区間)では人間は一定の印象を持つと仮定し、その区間をフレーズと定義する。そして、楽曲のフレーズ毎の印象を Russell の V-A 平面¹⁶⁾上に時系列順にプロットしたものを印象軌跡とする。V-A 平面は図1のように人間の感情状態を二次元平面で表現するモデルである。横軸は Valence 軸と呼ばれ快-不快を反映し、縦軸は Arousal 軸と呼ばれ興奮-弛緩を反映している。

2.2 本稿での処理

本研究の処理の概要を図2に示す。歌詞と音響信号が既知な入力楽曲に対して、歌詞からは歌詞印象軌跡を、音響信号からは音響印象軌跡をそれぞれ推定する。具体的には、フレーズに切り分けられた歌詞と音響信号から、歌詞印象と音響印象をそれぞれ V-A 平面上の点として推定し、それぞれの軌跡を歌詞・音響印象軌跡とする。最後に、楽曲全体の印象軌跡を得られた2種類の軌跡の組み合わせとして表現する。

歌詞印象軌跡の推定には印象表現語を事前知識として用いた制約付き PLSA¹⁷⁾を用いる。PLSA は、学習データとして用意された文書 (歌詞) 中の単語の共起関係を元に、文書の持つ潜在的なトピックを推定可能な確率的生成モデルである。単語の共起関係に基づいているため、教師ラベルを用意することなく大量の学習データを使用できるため、自然言語処理の分野で広く用いられている^{18),19)}。通常の PLSA では推定されるトピックが必ずしも印象を反映しないという問題があったが、本研究では印象表現語を事前知識として用いた MAP 推定によりトピックが印象を表すように制約することで、この問題を解決した。

音響印象軌跡の推定には、説明変数をフレーズの音響特徴量、目標変数を楽曲のフレーズが持つ V-A 平面上の座標とした重線形回帰分析を用いる。これにより、音響特徴量を入力とし、その音響特徴量が表す印象を V-A 平面上の点として推定することができる。事前に、

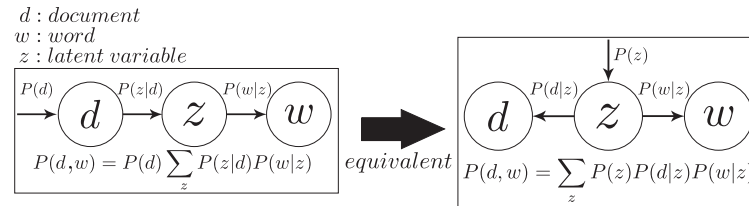


図3 PLSA のグラフィカルモデル

楽曲のフレーズとその印象を表す V-A 平面上の座標のペアからなる学習データを人手で用意し、各フレーズの音響信号から抽出された音響特徴量と対応する V-A 平面上の座標の関係を重線形回帰分析で分析する。実行時には、得られた重線形回帰分析パラメータを用いて入力楽曲の各フレーズの音響信号から 1 つの V-A 平面座標を計算し、それらの楽曲全体での軌跡を音響印象軌跡とする。以下、3 章では歌詞印象軌跡推定手法、4 章では音響印象軌跡推定手法を詳しく説明する。

3. 歌詞印象軌跡推定

歌詞の各フレーズが持つ単語の V-A 平面座標を推定してフレーズ毎の V-A 平面座標を求め、各座標を V-A 平面に順にプロットすれば歌詞印象軌跡が得られる。単語の V-A 平面座標の推定には、事前知識を用いた MAP 推定による PLSA¹⁷⁾ を用いる。なお、本手法を用いる前に歌詞をフレーズ毎に分割しておく必要がある。

3.1 確率的潜在意味解析

歌詞印象軌跡推定には PLSA を用いる。PLSA は文書のトピックによるクラスタリングや単語の潜在的意味抽出などに用いられる確率的生成モデルである。PLSA において、文書と単語の共起確率は文書のトピックを表現する潜在変数を導入した生成モデルで表現される (図 3)。文書 d から潜在変数として文書のトピック z が観測され、 z から文書中の単語 w が観測されるモデルを仮定し、 d と w の共起確率を以下のように定義して EM (Expectation-Maximization) アルゴリズムによりモデルパラメータを推定する。

$$P(d, w) = P(d) \sum_z P(z|d)P(w|z) = \sum_z P(z)P(d|z)P(w|z)$$

$P(w|z)$ は各トピックから各単語が観測される確率であり、 $P(w|z)$ が高い単語で z が表現するトピックが決定される。

3.2 モデルパラメータの MAP 推定

歌詞印象軌跡推定のためには、潜在変数 z に歌詞印象を表現させる必要があるが、通常の PLSA は文書と単語の共起確率のみに注目して文書のトピックや単語の潜在的意味を推定する手法であり、潜在変数が必ずしも印象を表現するとは限らない。そのため、本研究ではあらかじめ V-A 平面上の座標が求められている印象表現語 (happy, sad など) を事前知識として使用しモデルパラメータを MAP 推定することにより、各潜在変数に V-A 平面上的各象限を表現させるように制約をかける。

まず、潜在変数 z を以下のように定義する。

$$z \in \{V+A+, V+A-, V-A+, V-A-\}$$

すなわち、潜在変数 z の取り得る値に V-A 平面上的各象限を割り当てる。モデルパラメータの事前分布を、共役事前分布を用いて以下のように定義する。

$$\prod_w \prod_z P(w|z)^{\alpha_{w,z}-1}$$

印象表現語の $P(w|z)$ が持つ $\alpha_{w,z}$ に適切な値を与え、 $P(w|z)$ の推定値を大きくする。これにより、 z に V-A 平面上的各象限を表現させる事ができる。例えば $z_1 = V+A+$ であれば、V-A 平面上で V+A+ 象限に位置する happy, glad などに対して α_{w,z_1} を与える。

事前知識には ANEW²⁰⁾ と WordNet²¹⁾ を用いる。ANEW は英単語 1034 単語について V-A 平面上的座標を調査したデータであり、WordNet は同義語の集合を 1 ノードとして各ノードの関係 (下位語, 上位語, 対義語, 類義語など) をグラフにまとめたシソーラスである。 $\alpha_{w,z}$ の設定は以下を行う。

- (1) WordNet を用いて ANEW の同義語と類義語を探索し、ANEW を 1034 単語から 9757 単語に拡張する。
- (2) ANEW が存在する V-A 平面象限に対応する $\alpha_{w,z}$ を設定する。 $\alpha_{w,z}$ の値は、ANEW が原点にあれば 1、それ以外の場合は原点との距離に応じて値を大きくする。 $\alpha_{w,z}$ の最大値は 1.01 とする。この値は、予備実験により ANEW の V-A 平面座標が最も正確に推定できるよう定めた。

各フレーズの V-A 平面上的座標は、 $P(z|w)$ を用いて求めた各単語の V-A 空間座標を合計、正規化して求める。具体的には以下の式で求める。

$$V = \frac{1}{K} \sum_{k=1}^K ((P(V+A+|w_k) + P(V+A-|w_k)) - (P(V-A+|w_k) + P(V-A-|w_k)))$$

$$A = \frac{1}{K} \sum_{k=1}^K ((P(V+A+|w_k) + P(V-A+|w_k)) - (P(V+A-|w_k) + P(V-A-|w_k)))$$

K は各フレーズに含まれる単語数である。また、 $P(z|w)$ は $P(w|z)$, $P(z)$, $P(w)$ から計算可能である。最後に各フレーズの座標を V-A 平面上にプロットして、歌詞印象軌跡が得られる。

4. 音響印象軌跡推定

重線形回帰分析を用いて、音響信号から印象軌跡を V-A 平面座標の軌跡として推定する。まず、音響信号と対応する V-A 平面座標からなる学習データを用意し、音響信号から抽出された音響特徴量と V-A 平面座標との関係を学習する。その後、入力楽曲の音響特徴量から各フレーズの V-A 平面座標を推定し、各座標を V-A 平面にプロットすれば音響印象軌跡が得られる。

4.1 学習データ収集

学習データとして、各フレーズの V-A 平面座標とフレーズ切り替わり時刻を準備する必要がある。そのため、図 4 のようなグラフィカルユーザーインターフェースを開発した。このインターフェースは、Kim らが開発した時間変化する楽曲印象データを取得する為のインタラクティブゲーム²²⁾を参考に作成した。インターフェースのデザインには V-A 平面が反映されており、横軸が Valence 軸（快-不快）、縦軸が Arousal 軸（興奮-弛緩）を表す。ユーザは楽曲を聴取し、印象が切り替わったと判断した時点でインターフェースをクリックし、各フレーズの V-A 平面座標と楽曲中でフレーズが切り替わる時刻を記録する。記録した結果は行列の形で保存され、行列の各行にはフレーズ切り替わり時刻とフレーズの V-A 平面座標が格納される。

4.2 音響特徴量設計

本研究では、複数の音響特徴量から主成分分析を用いて有意な特徴量を選択し、入力する音響特徴量を設計する。これは、多岐に渡る音響特徴量が楽曲印象推定に用いられており、複数の特徴量を用いると印象推定精度が向上するという知見に基づく³⁾。

まず従来の楽曲印象推定手法⁹⁾⁻¹³⁾にならい、楽曲印象推定で頻繁に用いられる各フレーズの音響特徴量を抽出する。表 1 に抽出した特徴量と概要を示す。これらの特徴量を幅 32ms のフレームを使用しオーバーラップなしで抽出し、それらのフレーズ内平均と分散を 142 次元のベクトルとする。特徴量は 16kHz にダウンサンプリングしたモノラル音響信号から

表 1 フレーズから抽出する音響特徴量。これらの平均と分散からなる 142 次元のベクトルを、主成分分析により 22 次元に圧縮。

音響特徴量	概要
周波数スペクトル形状特徴	スペクトル重心、スペクトルフラックス、スペクトルローオフ、スペクトルフラットネスの 27 次元からなる。
メル周波数ケプストラム係数 (MFCC)	スペクトル包絡を表現する特徴量。音声認識や音楽音響信号のモデル化に使用される ²⁴⁾ 。本研究では 13 次元を使用する。
クロマベクトル	12 音名の各音名（ピッチクラス）の周波数のパワーを複数のオクターブに渡って加算した 12 次元の特徴量 ²⁵⁾ 。
線スペクトル対	線形予測係数と等価な周波数領域の係数で音声符号化で用いられる。本研究では 18 次元を使用する。
ゼロクロッシング	音響信号の時間波形がゼロを通過する回数を表す 1 次元の特徴量。

表 2 音響特徴量主成分の各クラス内での平均と分散、寄与率、および寄与した音響特徴量。

	第 1 主成分	第 2 主成分	第 3 主成分
寄与率	48.5%	10.6%	8.06%
寄与する特徴量	スペクトル重心分散 MFCC 平均 線スペクトル対平均	クロマベクトル分散 線スペクトル対分散	クロマベクトル分散 クロマベクトル平均 線スペクトル対分散

MARSYAS²³⁾を用いて抽出した。このベクトルの各次元を平均が 0、分散が 1 となるよう正規化し、主成分分析により累積寄与率 90%で 22 次元まで次元圧縮を行う。この 22 次元ベクトルをフレーズの特徴量とする。

表 2 に、圧縮後の寄与率が 8%以上の主成分と、それぞれに寄与する音響特徴量を示す。寄与する特徴量は、各主成分軸の絶対値の上位 5 つに含まれている特徴量を選択した。表 2 より、寄与率が高い特徴量にはクロマベクトル、MFCC、線スペクトル対、スペクトル重心が寄与している事がわかる。

4.3 重線形回帰分析

説明変数は各フレーズの音響特徴量、目標変数は各フレーズの V-A 平面座標として多項式基底を用いた重線形回帰分析を行い、音響特徴量から各フレーズの V-A 平面座標を推定する。回帰式は以下のとおりである。

$$V = \sum_{i=0}^{M-1} \sum_{j=1}^{22} v_{ij}(x_j)^i, A = \sum_{i=0}^{M-1} \sum_{j=1}^{22} a_{ij}(x_j)^i$$

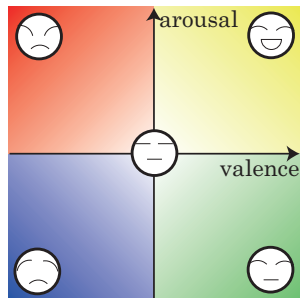


図 4 各フレーズの V-A 平面座標取得用 GUI. 横軸は Valence 軸, 縦軸は Arousal 軸を反映している.

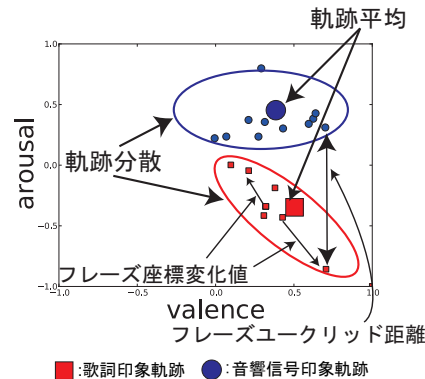


図 5 印象軌跡特徴

表 3 定義した印象軌跡特徴の詳細

印象軌跡特徴	詳細
軌跡平均	軌跡内の全フレーズ座標の平均.
軌跡分散	軌跡内の全フレーズ座標の共分散.
軌跡ユークリッド距離	歌詞 V-A 平面座標と音響信号 V-A 平面座標のユークリッド距離 (フレーズユークリッド距離) の軌跡内での合計.
フレーズ座標変化値分散	軌跡内でのフレーズ座標変化値の分散.

x_j は 22 次元の特徴量ベクトルの j 番目の要素, v_{ij} , a_{ij} は重線形回帰分析のパラメータ, M はパラメータ数である. 基底関数のパラメータ数 M は予備実験により最も予測誤差が小さくなる 3 を選択した. 学習データより推定したパラメータ v_{ij} , a_{ij} を用いて, 入力楽曲の各フレーズの音響信号から V-A 平面座標を求める. そして, 推定した各フレーズの V-A 平面座標の楽曲全体での軌跡として音響印象軌跡が得る.

5. 評価実験

本稿で開発した歌詞と音響信号の印象軌跡推定手法の妥当性を評価するため, 評価実験を行う. 具体的には, まず評価用楽曲群から推定した印象軌跡を軌跡の類似度に基づきクラスターリングすることで, 歌詞・音響印象軌跡の形状と配置場所に基づいて楽曲をクラス分けする. そして, 各クラスの音響特徴量の分散を比較して異なる傾向が見られる事を確かめる事で, 印象軌跡が楽曲印象の時間変化構造を表現できる事を確認する. また, 楽曲を表現する

ソーシャルタグをウェブ上から収集し, 印象軌跡とタグの間に矛盾が見られない事を確認することで, 提案法により適切な印象軌跡が推定出来る事を確かめる.

5.1 実験概要

データセットとして, “The Beatles” の楽曲 175 曲を使用する. これらの楽曲は音響信号と歌詞が既知であり, また和音系列正解データなどが公開されている楽曲である²⁶⁾. MIR 研究分野で広く用いられているため採用した. 実験は, leave-one-out 法を用いて行う. 具体的には, ある楽曲の印象軌跡を推定する際には, 残りの 174 曲を使用して歌詞, 音響印象軌跡推定法のモデルを学習する. この操作を全 175 曲に対して行うことで, すべての楽曲に対して歌詞, 音響印象軌跡を推定できる.

まず, 対象楽曲に対して学習データを準備する. 全曲について歌詞を読みながら音響信号を聞き歌詞をフレーズ毎に分割し, 4.1 節で述べたグラフィカルユーザインタフェースを用いてフレーズ切り替わり時刻と各フレーズの音響信号の印象を表す V-A 平面座標をアノテーションする. 以上の準備は, 一貫性を保つために本論文の主著者一人が実施した.

さらに, 対象楽曲を特徴付けるソーシャルタグを, ウェブ上から収集する. ソーシャルタグはウェブ上で不特定多数の人間が楽曲データ (印象, 作曲者, ジャンルなど) に応じて楽曲にタグ付けするものであり, 楽曲印象が反映されやすい. そのため, 楽曲印象推定研究でも頻繁に用いられている^{27),28)}. ソーシャルタグはソーシャルネットワーキングサービス last.fm (<http://www.last.fm/>) 上で 5 回以上タグ付けされている形容詞のみを収集する. これは, 印象表現語のほとんどが形容詞であるためである.

5.2 階層的クラスターリング

クラスターリングにはワード法による階層的クラスターリング手法を用いる. 楽曲印象時間構造の複雑度で分類されているかの判断は, 各クラス内楽曲の音響特徴量の比較で行う.

クラスターリングのための特徴量として, 歌詞印象軌跡と音響印象軌跡から図 5, 表 3 に示す特徴量を使用する. これらの特徴量は, 印象軌跡の 4.2 節の音響特徴量と区別するため, 本稿では印象軌跡特徴量と呼ぶ. この印象軌跡特徴量を用いてワード法を用いた階層的クラスターリングを行う.

5.3 各クラスが持つ音響特徴量の比較

クラスターリングで得られた各クラスの音響特徴量の分散を比較する. 楽曲印象の時間構造が複雑な曲 (すなわち印象軌跡の分散が大きいことが期待される曲) では, フレーズ毎に音響信号が大きく変化していると考えられる. そのため, 各クラスの個々の楽曲が持つ音響特徴量の分散を比較して分散の大小で楽曲が分類されている事を確認する事で, 印象軌跡が楽

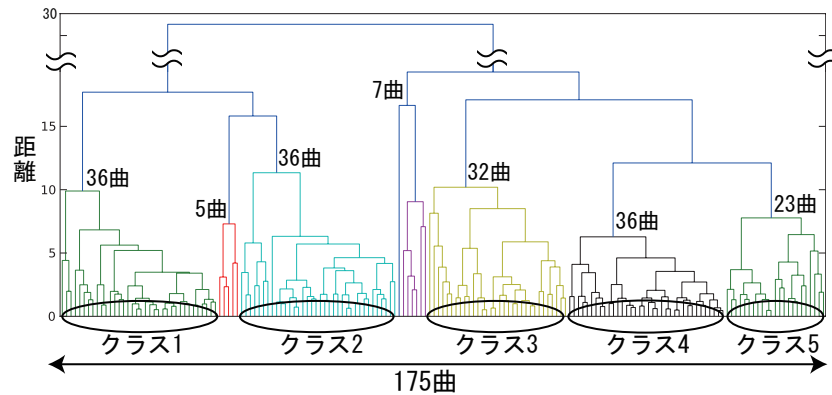


図 6 階層的クラスタリング結果

曲印象時間構造を表現可能かを検証できる。

5.4 各クラスのソーシャルタグと印象軌跡の比較

クラスタリングで得られたクラスの楽曲が共有するソーシャルタグと印象軌跡を比較する。ソーシャルタグはウェブ上で不特定多数の人間が楽曲データ（印象、作曲者、ジャンルなど）に応じて楽曲にタグ付けするものである。ソーシャルタグを用いた楽曲印象推定研究^{27),28)}も行われており、ソーシャルタグには楽曲印象が反映されやすい。そのため、ソーシャルタグの印象を表現する V-A 平面座標と推定された印象軌跡の距離を確認する事で、印象軌跡の妥当性が検証できる。

5.5 結 果

5.5.1 クラスタリング結果

図 6 は階層的クラスタリングで楽曲をクラスタリングしたデンドログラムである。横軸は実験に用いた全楽曲、縦軸は印象軌跡特徴の距離である。クラスタ間距離の最大値の 0.7 倍をスレッシュホールドとして色分けを行った。デンドログラムは 8 色に色分けされているが、赤、紫、青の楽曲は他の色が付けられている楽曲群の半数以下しか存在しない。よって、印象軌跡を用いたクラスタリングによって主要な 5 個のクラスが得られると判断する。

5.5.2 各クラス音響特徴量の分散の比較結果

フレーズ音響特徴量の楽曲内分散を計算し、各クラス内で平均したものを表 4 に示す。比較した特徴量は、4.2 節で示した寄与率 8% 以上の主成分である。表 4 より、クラス 1 では第 1 主成分の分散が 2 番目に大きく、第 2、第 3 主成分では分散が最大となっている。また

表 4 各クラスの音響特徴量の分散のクラス内平均。

	第 1 主成分	第 2 主成分	第 3 主成分
クラス 1	12.47	2.96	2.42
クラス 2	9.62	2.01	1.95
クラス 3	9.04	0.60	1.39
クラス 4	12.81	1.61	1.13
クラス 5	10.69	3.14	2.24

クラス 3 では第 1、第 2 主成分の分散が最小、第 3 主成分の分散が 2 番目に小さい。以上より、クラス 1 は楽曲印象の時間構造が複雑なクラス、クラス 3 は楽曲印象の時間構造が単純なクラスとして分類された事が予想される。

5.5.3 各クラスのソーシャルタグと印象軌跡の比較結果

表 5 に各クラス内で複数の曲にタグ付けされているソーシャルタグを示す。なお、表 5 の右端の列には実験に用いた楽曲全体で登場曲数が最も多いタグの上位 5 位を示す。「classic, psychedelic, beautiful, happy, mellow」は実験に用いた楽曲全体に広く付与されているので、実験データ全体の傾向であると判断できる。そのため、これらのソーシャルタグは無視した。図 7 は、各クラス内楽曲の印象軌跡の分布とソーシャルタグが表現する印象を表現する V-A 平面座標をプロットしたものである。ソーシャルタグの V-A 平面座標は ANEW, WordNet より求めた。ソーシャルタグが ANEW 内に含まれている場合は ANEW の V-A 平面座標を用いた。タグが ANEW に含まれていない場合は WordNet で ANEW に含まれる類義語を探索し、ANEW が見つかった場合はその V-A 平面座標を用いた。ANEW に含まれる類義語が見つからない場合は、そのソーシャルタグは無視した。

クラス 1 では V-A 平面にマッピングした 6 個のソーシャルタグのうち、「melancholic, relaxed」は印象軌跡分布から遠いものの、「sweet, loved, romantic, cute」が印象軌跡分布に近い。つまり、半分以上のタグが表す印象が印象軌跡の表現する印象と合致したと言える。その他、クラス 3 から 5 でも、半数以上のソーシャルタグの表す印象が、音響信号または歌詞印象軌跡分布と合致している。一方、クラス 2 では 4 個のソーシャルタグのうち、「sweet」のみが印象軌跡分布に近く、その他のソーシャルタグは推定された印象軌跡分布と合致しなかった。以上より 5 クラス中 4 のクラスで、過半数のソーシャルタグの表現する印象が印象軌跡と合致した。この結果より、本手法で適切な印象軌跡が推定されたと考えられる。

また、各クラスの音響信号および歌詞印象軌跡分布の位置を観察すると、クラス 1 と 2 で

表 5 各クラスが共有しているソーシャルタグ。タグは last.fm 上で収集した。右端の列はタグ付けされている曲数が多いタグの上位 5 位。

クラス 1	クラス 2	クラス 3	クラス 4	クラス 5	top 5 tags
sweet	melancholic	melancholic	romantic	romantic	classic
experimental	sad	relaxed	sad	experimental	psychedelic
amazing	lovely	funny	sweet	melancholic	beautiful
relaxed	trippy	romantic	chill		happy
chill	romantic	chill	trippy		mellow
catchy	sweet	sweet			
loved		lovely			
romantic		cute			
trippy					
melancholic					
cool					
cute					

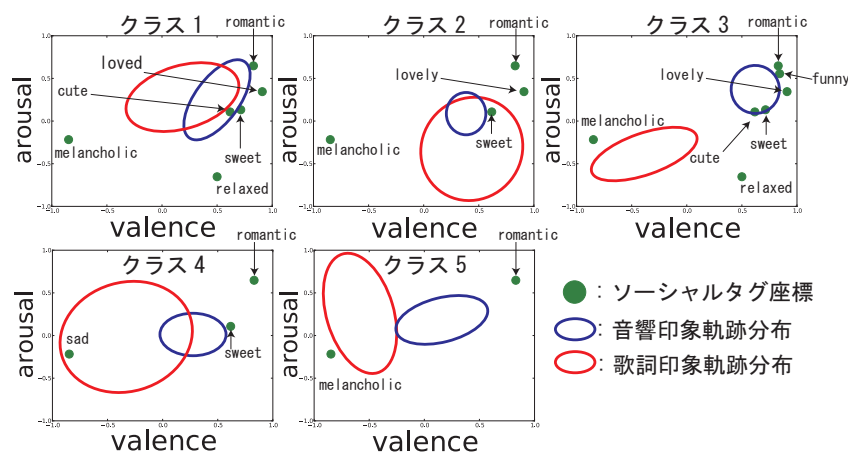


図 7 各クラスの印象軌跡分布とマッピングしたソーシャルタグ

は歌詞と音響信号の印象軌跡分布が大きく重なっているため、歌詞と音響信号の印象が類似している事が予想される。一方、クラス 3 では歌詞と音響信号の印象軌跡分布が大きく離れており、歌詞と音響信号の印象が異なっている事が予想される。

6. おわりに

本稿では歌詞と音響信号を用いた歌詞・音響印象軌跡推定手法を提案した。楽曲印象を印象軌跡としてとらえることで、歌詞と音響信号の印象の類似度合いや楽曲印象の時間的構造が表現できる。評価実験として推定された印象軌跡を用いたクラスタリング結果の音響特徴量を比較し、楽曲印象の時間構造の複雑さで楽曲が分類可能である事が確認された。また、各クラスの印象軌跡分布とソーシャルタグの比較から、推定された印象軌跡は適切である事が明らかとなった。今後は推定された印象軌跡の妥当性を検証するための被験者実験、印象軌跡推定結果と従来手法で推定された楽曲印象推定結果の比較、印象軌跡を用いた音楽情報処理システムの実装と評価などを行う予定である。

参考文献

- 1) James Bergstra, Michael Mandel, and Douglas Eck: Scalable Genre and Tag Prediction with Spectral Covariance, ISMIR2010, pp.507-512, 2010.
- 2) Cheng Yang: Music Database Retrieval Based on Spectral Similarity, ISMIR2001, pp.37-38, 2001.
- 3) Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon G. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull: Music Emotion Recognition: A State of the Art Review, ISMIR2010, pp.255-266, 2010.
- 4) Yajie Hu, Xiaou Chen and Deshun Yang: Lyric-Based Song Emotion Detection With Affective Lexicon and Fuzzy Clustering Method, ISMIR2009, pp.123-128, 2009.
- 5) Menno van Zaanen and Pieter Kanter: Automatic Mood Classification Using TF*IDF Based On Lyrics, ISMIR2010, pp.75-80, 2010.
- 6) Tao Li and Mitsunori Ogihara: Detecting Emotion in Music, ISMIR2003, pp.239-240, 2003.
- 7) Janto Skowronek, Martin McKinney and Steven van de Par: A Demonstrator for Automatic Music Mood Estimation, ISMIR2007, pp.345-346, 2007.
- 8) Tuomas Eerola, Olivier Lartillot and Petri Toivainen: Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models, ISMIR2009, pp.621-626, 2009.
- 9) Dan Yang and Wonsook Lee: Disambiguating music emotion using software agents, ISMIR2004, pp.52-58, 2004.
- 10) Cyril Laurier, Jens Grivolla, and Perfecto Herrera, Multimodal Music Mood Classification Using Audio and Lyrics, ICMLA 2010, pp.688-693, 2008.

- 11) Xiao Hu, J. Stephen Downie, and Andreas F. Ehmann, Lyric Text Mining in Music Mood Classification, ISMIR2009, pp.411-416, 2009
- 12) Erik M. Schmidt and Youngmoo E. Kim: Prediction of Time-varying Musical Mood Distributions from Audio, ISMIR2010, pp.465-470, 2010.
- 13) Erik M. Schmidt, Douglas Turnbull, and Youngmoo. E. Kim, Feature Selection for Content-based, Time-varying Musical Emotion Regression, ACM SIGMM MIR 2011, pp.267-274, 2010.
- 14) 草間かおり, 伊藤貴之: MusCat:楽曲の印象表現に基づいた一覧表示の一手法, 情報処理学会研究報告, Vol.2009-MUS-81 No.19, 2009.
- 15) 三好 真人, 柘植 寛, Choge Kipsang Hillary, 尾山 匡浩, 伊藤 桃代, 福見 稔: 音楽検索のための楽曲印象値の自動付与手法, 情報処理学会研究報告, Vol. 2011-MUS-89 No. 23, 2011.
- 16) James A. Russell: A Ciucumplex Model of Affect, Journal of Personality and Social Psychology, Vol.39, No.6, pp.1161-1178, 1980.
- 17) Thomas Hofmann: Probabilistic Latent Semantic Analysis, UAI99, pp.289-296, 1999.
- 18) Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu: Topic-bridged PLSA for Cross-domain Text Classification, SIGIR '08, pp.627-634, 2008.
- 19) Yuya Akita and Tatsuya Kawahara: Language Model Adaptation based on PLSA of Topics and Speakers, ICSLP2004, pp.602-605, 2004.
- 20) Margaret M. Bradley and Peter J. Lang: Affective Norms for English Words (ANEW): Instruction Manual and Affective Rating, Technical Report, C-1, The Center for Research in Psychophysiology, University of Florida, 1999.
- 21) George A. Miller: WordNet: A Lexical Database for English, Communications of the ACM, Vol.38, Issue.11, pp.39-41, 1995.
- 22) Youngmoo E. Kim, Erik M. Schmidt, and Lloyd Emelle: MoodSwings: A Collaborative Game for Music Mood Label Collection, ISMIR2008, pp.231-236, 2008.
- 23) George Tzanetakis and Perry Cook: MARSYAS: A Framework for Audio Analysis, Organised Sound, Vol.4, Issue.3, pp.169-175, 2000.
- 24) Beth Logan: Mel Frequency Cepstral Coefficients for Music Modeling, ISMIR2000, 11p., 2000.
- 25) Mark A. Bartsch and Gregory H. Wakefield: To catch a chorus: Usingchroma-based Representations for Audio Thumbnailing, WASPAA'01, pp.15-18, 2001.
- 26) Christopher Harte, Mark Sandler, Samer Abdallah, and Emilia Gómez: Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations, ISMIR2005, pp.66-71, 2005.
- 27) Kerstin Bischoff, Claudiu S. Firan: Music Mood and Theme Classification - A Hybrid Approach, ISMIR2009, pp.657-662, 2009.
- 28) Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu: How Do You Feel about 'Dancing Queen'? : Deriving Mood & Theme Annotations from User Tags, JCDL'09, pp.285-294, 2009.