1S-7

楽曲印象軌跡に基づく楽曲検索システムの実装と評価

西川 直毅 † 糸山 克寿 † 藤原 弘将 ‡ 後藤 真孝 ‡ 尾形 哲也 † 奥乃 博 †

† 京都大学大学院 情報学研究科 知能情報学専攻 ‡ 産業技術総合研究所

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ている．MIR システムを構成する上で人間が音楽をどのように知覚するかを考える事は非常に重要である．人間は音楽を印象の表現として知覚する [1] 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て計算機上で表現し，その印象に基づいて楽曲を検索する機能が実現できれば，MIR は人間の感性に適した検索結果を提供する事が可能となる．

本稿では，楽曲印象の類似度に基づく楽曲検索システムについて述べる．システムへの入力は楽曲音響信号と対応する歌詞テキストである．まずクエリ楽曲と、検索対象楽曲データベース中の各楽曲から楽曲印象を推定する．そして，推定された楽曲印象同士の類似度を計算し，入力楽曲と類似した印象を持つ楽曲を出力する．

楽曲印象の推定には，西川ら [2] によって提案された歌詞-音響信号印象軌跡推定手法を用いる．本手法は楽曲全体の印象として，歌詞と音響信号がそれぞれ持つ時間変化印象を組み合わせて使用する．これは，楽曲印象を推定する上で歌詞と音響信号を両方用いる事，印象の時間変化を考慮する事が重要であるとの考えに基づいている．例えば，明るい曲調だが歌詞が暗い場合など歌詞と音響信号で受ける印象が異なる楽曲，サビや A メロなどの区間によって曲調が大きく異なる楽曲，歌詞が物語のような時間構造を持つ楽曲などの印象は，先述の 2 点を考慮しないと正確な楽曲印象は表現できない．印象軌跡同士の類似度計算には DP マッチング [3] を使用する．

歌詞-音響信号印象軌跡の有効性を確かめる評価実験として，入力に歌詞-音響信号印象軌跡の両者を使用した場合，一方のみを使用した場合，印象の時間変化を表現しない印象軌跡の平均を使用した場合での検索結果の比較を行った．比較には印象タグ一致度を用いた．印象タグは，ウェブ上で不特定多数の人間が楽曲印象に応じて楽曲に付与するタグであり，楽曲印象の正解ラベルとして使用できる．

2 歌詞-音響信号印象軌跡に基づく楽曲検索システム

はじめに，歌詞-音響信号印象軌跡と本稿で述べる検索システムについて説明する．歌詞-音響信号印象軌跡の例を図 1 の左に示す．人間の感情状態を 2 次元平面上の座標で表現する V-A 平面 [4] 上に，歌詞と音響信号の印象の時系列変化がプロットされている．V-A 平面の縦軸は興奮-弛緩，横軸は快-不快を表現する．軌跡の各点は楽曲フレーズが持つ印象を表す．楽曲フレーズとは，ある程度一定の印象を持つと判断された楽曲内の区間である．

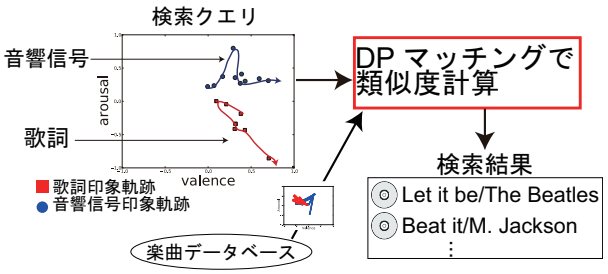

図 1: 印象軌跡の例と印象軌跡に基づく楽曲検索システム

印象軌跡推定手法 [2] では，事前に人手でフレーズ分割された楽曲を使用して印象軌跡を推定する．

図 1 は検索システムの概念図である．システムは予め，データベース中の楽曲の歌詞-音響信号印象軌跡を推定して保管しておく．システムに検索クエリとなる楽曲を入力すると，入力された楽曲の歌詞と音響信号から歌詞-音響信号印象軌跡が推定される．その後，DP マッチング [3] を使用してデータベース中の楽曲と入力楽曲の歌詞印象軌跡類似度，音響信号印象軌跡類似度をそれぞれ計算する．2 つの類似度の合計を楽曲類似度とし，検索結果として楽曲類似度が高い順にリスト形式で楽曲が出力される．

3 印象軌跡推定手法

本稿で使用した印象軌跡推定手法の概要について説明する．なお第 2 章で述べたように，本手法を適用する楽曲は事前にフレーズ分割されている事を前提とする．

3.1 歌詞印象軌跡推定: MAP 推定による PLSA

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 MAP 推定による PLSA [5] を用いる．歌詞フレーズ d から潜在変数として歌詞印象 z が観測され，z から歌詞中の単語 w が観測されると仮定し，d と w の共起確率を定義してモデルパラメータを推定し，歌詞印象を分析する．歌詞の各フレーズの V-A 座標は，$P(z|w)$ から求めた各単語の V-A 座標を正規化して求める．

通常の PLSA で推定される z は必ずしも印象を反映しない．本手法ではあらかじめ V-A 平面上の座標が明らかな印象表現語を事前知識としてモデルパラメータを MAP 推定してこの問題を解決する．具体的には以下のように各 z に V-A 平面上の各象限を表現させる制約をかける．

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

事前知識には WordNet[6] を用いて拡張した ANEW[7] を使用する．ANEW は英単語 1034 単語について V-A 平面上の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．歌詞中の単語のうち，拡張 ANEW に含まれる単語の $P(w|z)$ の推定値を大きくするように事前分布のパラメータを与える．例えば，V+A+ 象限に存在する happy に対しては，$P(\mathrm{happy}|\mathrm{V+A+})$ の推定値を大きく，

Implementation and Evaluation of Music Retrieval System based on Mood Trajectories: *Naoki Nishikawa (Kyoto Univ.), Katsutoshi Itoyama (Kyoto Univ.), Hiromasa Fujihara (AIST), Masataka Goto (AIST), Tetsuya Ogata (Kyoto Univ.), and Hiroshi G. Okuno (Kyoto Univ.)*

V−A− 象限に存在する sad に対しては，$P(\text{sad}|\text{V}-\text{A}-)$ の推定値を大きくするようにパラメータを与える．

3.2 音響信号印象軌跡推定: 重線形回帰分析

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，各フレーズの V-A 平面座標とフレーズ切り替わり時刻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分析する．使用する音響特徴量は MFCC (13 次元)，クロマベクトル (12 次元)，スペクトル形状特徴 (27 次元)，線スペクトル対 (18 次元)，ゼロクロス (1 次元) である．実行時には，得られた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．

4 評価実験

歌詞-音響信号印象軌跡に基づく楽曲検索システムの有効性を評価する為の実験を行った．具体的には，楽曲に付与されている印象タグを正解ラベルとして検索結果の評価を行った．印象タグとは，ウェブ上で不特定多数の人間が楽曲データ (印象やジャンルなど) に応じて付与したソーシャルタグのうち，楽曲印象を反映していると考えられるものを指す．本稿では Last.fm (`http://www.last.fm/`) で 5 回以上付与されているソーシャルタグから，3.1 節で述べた拡張 ANEW に含まれるものを印象タグとした．2 曲間の印象タグ一致度を以下のように定義する．

$$\frac{2\times(\text{tags}(A\cap B)/\text{tags}(A))\times(\text{tags}(A\cap B)/\text{tags}(B))}{(\text{tags}(A\cap B)/\text{tags}(A))+(\text{tags}(A\cap B)/\text{tags}(B))}$$

$\text{tags}(A\cap B)$ は曲 A と曲 B で共有されているタグ数，$\text{tags}(A)$ は曲 A，$\text{tags}(B)$ は曲 B が持つタグ数である．この式は 2 曲間で共有されているタグ数を各曲が持つタグ数で割った数値の調和平均と等価である．例えば $\text{tags}(A)=10$, $\text{tags}(B)=15$, $\text{tags}(A\cap B)=5$ の時は印象タグ一致度が 0.4 となる．本稿では入力楽曲とデータベース内の全楽曲との印象タグ一致度を求め，印象タグ一致度が閾値以上の楽曲を正解楽曲として検索結果を F 値で評価した．本稿では閾値を 0.21 とした．これは実験に使用する楽曲を使用して求めた全組み合わせの印象タグ一致度の平均の 2 倍に相当する．また F 値評価を行うために，システムには入力楽曲とデータベース内楽曲の平均楽曲類似度の 2 倍以上の類似度を持つ楽曲を検索結果として出力させた．

実験では，入力に歌詞-音響信号印象軌跡を使用した検索結果，歌詞-音響信号印象軌跡の一方のみを使用した結果，歌詞-音響信号印象軌跡平均 (印象の時間変化を考慮しない表現) を使用した結果を比較した．入力に印象軌跡平均を使用する場合は，入力楽曲とデータベース内楽曲が持つ歌詞-音響信号印象軌跡平均の V-A 平面上ユークリッド距離を計算し，2 つの距離の合計を楽曲類似度として検索を行った．使用した楽曲群は Last.Fm's Best of 2010 (`http://www.last.fm/bestof/2010/about`) から選択した 100 曲である．印象タグ平均は 5.2 個，最大タグ数は 18 個だった．全曲の歌詞に MAP-PLSA を適用して全曲の歌詞印象軌跡を推定し，leave-one-out 法で全曲の音響信号印象軌跡を推定する．なお，音響信号印象軌跡推定に使用する学習データ収集は 3 人の大学生に行わせた．全曲の歌詞-音響信号印象軌跡を推定した後，100 曲中のうち 1 曲を検索クエリとし，残りの 99 曲をシステムが持つ楽曲データベースとして検索する．検索は全 100 曲をクエリとして 100 回行った．なおこの実験方法では，検索対象となる楽曲群とデータベース内楽曲の類似度計算に両方の学習データが必要となり実際的ではないが，本稿の目的は印象軌跡の有効性の確認なので大きな問題ではない．

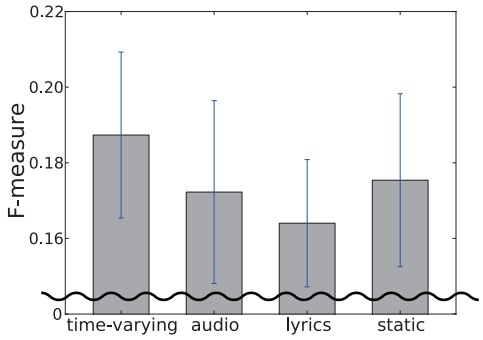

図 2: システムへの入力と平均 F 値．青線は分散を表す．time-varying は歌詞-音響信号印象軌跡を使用した結果，audio，lyrics は歌詞-音響信号印象軌跡の一方のみを使用した結果，static は印象軌跡平均を使用した結果を表す．

図 2 に，システムへの入力と 100 回の検索結果の平均 F 値を示す．歌詞-音響信号印象軌跡の一方のみを使用した場合や印象軌跡の平均を使用した場合よりも，歌詞-音響信号印象軌跡を使用した場合の方が F 値が高い．この結果から，「歌詞と音響信号がそれぞれ持つ時間変化印象」を表現する歌詞-音響信号印象軌跡によって楽曲検索システムはより人間の感性に適した検索結果を提供出来る事が明らかとなった．

5 おわりに

本稿では，歌詞-音響信号印象軌跡に基づく楽曲検索システムについて述べた．歌詞-音響信号印象軌跡は「歌詞と音響信号それぞれが持つ印象」と「楽曲印象の時間変化性」を同時に表現できる．評価実験から，歌詞-音響信号印象軌跡によってより人間の感性に適した楽曲検索が可能となる事が明らかとなった．今後は GUI インターフェースを実装し，被験者実験による検索システムの評価を行う予定である．なお，本研究は科研費 (S)，GCOE の支援を受けた．

参考文献

[1] Kim, Y. E. *et al*., "Music Emotion Recognition: A State of the Art Review", *ISMIR2010*, pp. 255–266, 2010.
[2] Nishikawa, N. *et al*., "A Musical Mood Trajectory Estimation Method Using Lyrics and Acoustic Features", *ACM MIRUM'11*, pp. 51–56, 2011.
[3] 迫江他, "動的計画法を利用した音声の時間正規化に基づく連続単語認識", 日本音響学会誌, Vol. 27, No. 9, pp. 483–490, 1971.
[4] Russell, J. A., "A Ciucumplex Model of Affect", *Journal of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161–1178, 1980.
[5] Hofmann, T., "Probabilistic Latent Semantic Analysis", *UAI99*, pp. 289–296, 1999.
[6] Miller, G. A., "WordNet: A Lexical Database for English", *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.
[7] Bradley, M. M. *et al*., "Affective Norms for English Words (ANEW)": Instruction Manual and Affective Rating, Technical report, The Center for Research in Psychophysiology, University of Florida, 1999.