5R-3

歌詞と音響特徴量を用いた楽曲の印象軌跡推定

西川 直毅 † 糸山 克寿 † 藤原 弘将 ‡ 後藤 真孝 ‡ 高橋 徹 † 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 (AIST)

1. はじめに

楽曲を視聴した時に引き起こされる感情状態は一般的に楽曲の印象と呼ばれる．楽曲の印象は時間によって変化するが，ある一定の区間 (一般に A メロ，B メロ，サビなどと呼ばれる数十秒程度の区間) では大きく変化しないと考えられる．本稿では，楽曲中で印象が大きく変化しない区間を「フレーズ」，楽曲のフレーズ毎の印象を時系列順にプロットしたもの「印象軌跡」と定義する．従来の楽曲印象推定研究 [1] は主に機械学習を用いた楽曲単位での印象推定を対象としているが，印象軌跡によって曲全体の印象は左右されると考えられる．つまり，従来法では曲全体の印象を正確に推定する事は難しい．

本稿は楽曲の印象軌跡推定手法について議論する．印象軌跡推定が可能になれば，印象類似楽曲検索の精度向上，1 曲に複数の印象タグ付けをすること等が実現でき，音楽情報検索の新しいアプローチになると期待される．

本稿では推定する印象を具体的に定義し，歌詞を文書，歌詞が表現する印象を文書のトピックとした確率的潜在意味解析 (pLSA) [2] を行いフレーズ印象を推定する．pLSA で得られるトピックは必ずしも印象には対応しないものの，各トピックに代表的な印象表現語を事前知識として割り当てて最大事後確率 (MAP) 推定する事で，トピックと印象が 1 対 1 対応する．MAP-pLSA により歌詞中の単語から印象が観察される確率を推定し，フレーズ中の単語全てについて合計したものをフレーズ印象と定義する．次に，フレーズ印象とフレーズ音響特徴との相関を多項式回帰を用いて学習する．本稿ではフレーズ印象が音響特徴量で予測可能と仮定している．

未知楽曲の音響特徴量を用いた印象軌跡推定実験では，歌詞から推定した印象軌跡と音響特徴量から推定した印象軌跡が異なる例が存在した．2 つの軌跡はそれぞれ歌詞，音響信号の印象を反映した結果であり，歌詞のみでは推定できない印象軌跡及び，逆に音響特徴量のみでは推定できない印象軌跡の存在が示唆された．

2. 印象の定義

本稿で推定する印象は Russell の円環モデル [5] を用いて定義する．このモデルは，図 1 のように感情状態を 2 次元平面で表現する．*Valence* 軸は快-不快，*Arousal* 軸は興奮-弛緩を表現する．円環モデルより，印象は V+A+ (快，興奮)，V+A− (快，弛緩)，V−A+ (不快，興奮)，V−A− (不快，弛緩) の 4 つ，つまり推定すべきフレーズ印象は 4 次元ベクトルとなる．

3. 印象軌跡推定

フレーズ印象推定に用いるフレーズの歌詞を得るために，次の 2 つの前処理を行う．

1. 楽曲の印象変化時間の取得
2. 歌詞と音響信号の同期

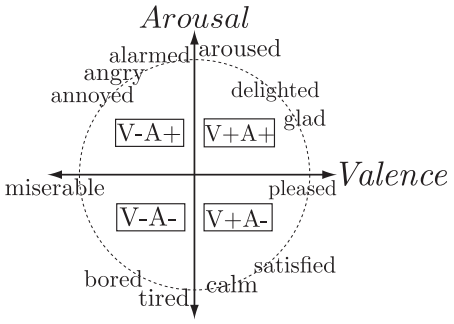

図 1: Russell の円環モデル

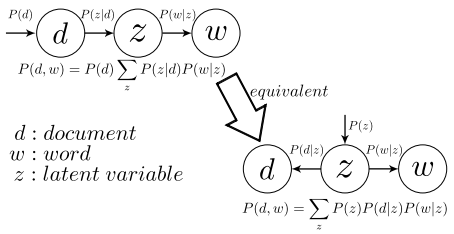

図 2: pLSA

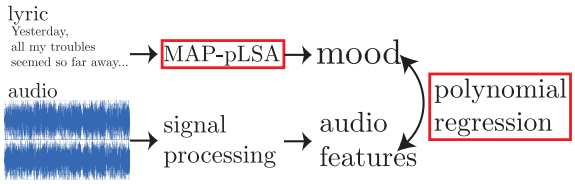

図 3: 印象軌跡推定概要．赤で囲われている MAP-pLSA，音響特徴と印象の相関学習について本稿で主に議論する．

印象変化時間を自動で取得するシステムは存在しないので，本稿では使用する楽曲を全て聞いてフレーズ印象を判断し，印象が変化した時間を手動で記録する．歌詞と音響信号の同期は手動で行う．なお，歌詞と音響信号の同期システムには，藤原らの Lyric Synchroniser [3] などが存在する．次に図 3 に示すように，以下の 2 処理を行う．

1. MAP-pLSA による歌詞からのフレーズ印象推定
2. 音響特徴と印象の相関を学習

3.1 ANEW, WordNet を用いた MAP-pLSA によるフレーズ印象推定

本稿では pLSA を用いて歌詞から各フレーズ印象を推定する．pLSA は，文書 d からトピック z が生成されトピック z から単語 w が生成されるという確率モデル (図 2) に基づいて，文書 d と単語 w の共起確率 $P(d,w)$ から文書や単語とトピックとの関係を分析する手法である．文書 d を歌詞，単語 w を歌詞中の単語，トピック z を楽曲の印象とみなすことで，印象に基づく歌詞と単語のクラスタリングや印象を代表する単語の抽出が可能となる．pLSA，モデルの対数尤度関数はそれぞれ以下の式で表現される．

$$P(d,w) = P(d)\sum_z P(z|d)P(w|z) = \sum_z P(z)P(d|z)P(w|z) \quad (1)$$

Musical Mood Trajectory Estimation using Lyrics and Audio Features: *Naoki Nishikawa, Katsutoshi Itoyama (Kyoto Univ), Hiromasa Fujihara, Masataka Goto (AIST), Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

$$\log L = \sum_d \sum_w n(d,w) \log(P(d,w)) \qquad (2)$$

$n(d,w)$ は単語文書共起頻度であり，tf-idf を用いて求める．対数尤度関数を最大化することで，共起頻度をもっともよく表現する $P(z)$, $P(d|z)$, $P(w|z)$ を求める．z は潜在変数であるため，対数尤度関数の最大化には EM アルゴリズムを用いる．

$P(z|w)$ は単語 w からトピック z が観測される確率であり，$P(w|z)$ から計算される．文書に含まれる各単語に対する $P(z|w)$ を合計して正規化することで，未知の文書に対してもトピックを推定することができる．

「明るい」「悲しい」といった印象の多くは，「happy」「sorrow」のような，その印象と関連の深い単語の集合として表現される．一方，pLSA は歌詞と単語の共起関係のみに基づく分析手法であるため，得られたトピックは必ずしも印象を表現するわけではない．すなわち特定の印象とは結びつかない単語 (助詞や接続詞など) で代表されるトピックが得られてしまう場合がある．このような不適切な場合を回避するために，各トピックにあらかじめ代表的な印象表現語を事前知識として割り当て，事前知識と共起関係の両方をもっともよく表現するように各確率を MAP 推定する．なお，潜在変数に 2 章で定義した 4 印象を割り当てる．

MAP 推定に必要な対数事前分布を以下で定義する．

$$\log P(\theta) \propto \sum_w \sum_z (\alpha_{w,z} - 1) \log P(w|z) \qquad (3)$$

印象表現語の $P(w|z)$ に対応する $\alpha_{w,z}$ を導入し，印象表現語の $P(w|z)$ 推定結果を大きくする．$\alpha_{w,z}$ の値の設定には ANEW [6] と WordNet [7] を用いる．ここで，ANEW は 1034 の英単語について Russell の円環モデル上の座標を調査したデータであり，WordNet は同義語の集合を一つのノードとし，各ノードの関係 (下位語，上位語，対義語，類義語等) をグラフにまとめたシソーラスである．WordNet を用いて ANEW の同義語を探索して 1034 単語から 9757 単語に拡張し，全単語に円環モデルの原点からの距離に比例して $\alpha_{w,z}$ を 1 から 5 の範囲で与える．拡張した ANEW に含まれない単語の $\alpha_{w,z}$ は設定しない．$P(z)$, $P(d|z)$ に対しては無情報事前分布を与える．以上によって求められた $P(z|w)$ をフレーズに含まれる単語全てについて合計，正規化し，4 次元のフレーズ印象を求める．

3.2 フレーズ印象と音響特徴量の相関

フレーズの音響特徴量は以下の定義とする．まず信号処理ツールの MARSYAS [4] を用いて取得した楽曲全体の特徴量からフレーズ部分を切り出す．切り出した特徴量のうち，楽曲の印象推定で主に用いられる [1] メル周波数ケプストラム係数 (MFCC) 13 次元，クロマベクトル 12 次元，スペクトルフラックス，ゼロクロス，スペクトル重心各 1 次元，これらのフレーズ内平均と分散の 56 次元ベクトルをフレーズ音響特徴量とする．

本稿では，歌詞から得たフレーズ印象は音響特徴量で予測可能と仮定する．この仮定に基づき，各フレーズの音響特徴量を説明変数，各フレーズ印象を目標変数とした多項式回帰を用いてフレーズの音響特徴量からフレーズ印象を予測する．多項式のパラメータ数は 10-fold cross validation により 7 とした．次に，回帰に用いなかったフレーズの音響特徴量を回帰式に入力してフレーズ印象を予測し，印象軌跡を推定する．

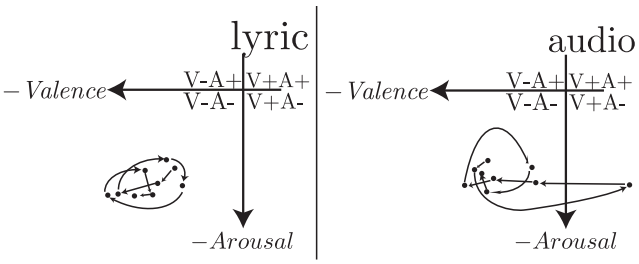

図 4: 歌詞から推定した印象軌跡と音響特徴量から推定した印象軌跡が異なる例．黒点が各フレーズを表す．左が歌詞から推定した印象軌跡，右が音響特徴量から推定した印象軌跡である．軌跡の変化は，歌詞のみでは表現出来ない印象及び音響特徴量のみでは表現できない印象の存在を示唆する．

4. 印象軌跡追跡実験

MAP-pLSA に用いたビートルズの楽曲 179 曲を用いて手動でフレーズ分割，歌詞同期を行ったのち，169 曲のフレーズの音響特徴量を説明変数，印象を目標変数とした回帰問題を解いた．その後回帰に用いなかった 10 曲のフレーズの音響特徴を用いて印象軌跡を推定すると，10 曲中 1 曲で印象軌跡が大きく変化した．結果を図 4 に示す．なおこの楽曲のタイトルは「She said, she said」である．

この曲はスローテンポで明るい曲調だが，歌詞には bad, sad, dead, no などの単語が多用されており，ネガティブな内容が歌われている．推定結果を見ると，歌詞から推定した印象軌跡は V−A− (不快，弛緩) 領域に偏っているのに対し，音響特徴量から推定した軌跡は 3 フレーズ目が V+A− (快，弛緩) 領域に移動している．また軌跡全体をみると，音響特徴量から推定した印象軌跡は全体的に V+ (快) の方向に移動している．以上の 2 点より，音響特徴量から推定した印象軌跡は歌詞，音響特徴量それぞれの印象が反映されている推定結果であると考えられる．またこの結果は，歌詞のみでは推定できない印象軌跡及び音響特徴量のみでは推定できない印象軌跡の存在を示している．つまり，印象軌跡はこのように歌詞と音響特徴量のフラットな統合ではなく，両者のダイナミクスをとらえた階層的な表現とする必要があると考えられる．

5. おわりに

本稿では，MAP-pLSA によって得たフレーズ印象と音響特徴量の多項式回帰を用いた楽曲の印象軌跡推定手法について議論した．今後の課題としては，被験者実験を通した印象軌跡推定結果の評価と精度向上及び階層的な印象軌跡表現法の検討が挙げられる．

謝辞 本研究は科研費 (S), JST CrestMuse, GCOE の支援を受けた．

参考文献

[1] Youngmoo E. Kim et al.: “Music Emotion Recognition: A State of the Art Review”, ISMIR, 2010.
[2] Thomas Hofmann: “Probabilistic Latent Semantic Analysis”, UAI, 1999.
[3] Hiromasa Fujihara et al.: “Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals” ISM, 2006.
[4] G. Tzanetakis et al.: “MARSYAS: A Framework for Audio Analysis”, Organised Sound, Vol.4, Issue.3, pp.169–175, 2000.
[5] James A. Russell: “A Ciucumplex Model of Affect”, Journal of Personality and Social Psychology, Vol.39, No.6, pp.1161–1178, 1980.
[6] Margaret M. Bradley et al.: “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.
[7] George A. Miller: “WordNet: a lexical database for English”, Communications of the ACM, Vol.38, Issue.11, pp.39–41, 1995.

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.