

グラフカーネルフィルタをもちいた楽曲の画一な印象の獲得

Acquisition of uniform impression of music using graph kernel filter

錠 尚史^{*1}

Hisashi IKARI

^{*1}株式会社ソケット

Sockets Co., Ltd.

The impression of a song varies from person to person. The sensitivity that depends on this individual is greatly different among individuals and it is difficult to quantify. In addition, because sensitivity changes with time, fixed point observation is also necessary, and it takes much time. In this case, focusing on the Graph laplacian which has the property of Fourier transform among the Deep learning attracting attention in recent years, we will clarify the time shortening by computer experiment and the frequency feature which affects the sensitivity of music.

1. はじめに

楽曲の印象は人により様々である。この個人に依存する感性は個人差があり、楽曲の印象の研究は心理学実験 [Hevner, 1936] を始めとする統計数理による検定に基づく研究が主立つ。この際、統計をもちいるために必要なサンプルの数は多程、楽曲の印象の真の分布に近づく [Clopper and Pearson, 1934]。しかしながら、大規模の心理学実験には時間と人員を要し、また、感性は時間経過に伴い変化するために、時間変化を考慮した定点観測も必要であり、多くの時間を要する。

近年、深層学習の一般物体の認識の精度が注目を浴びている [Krizhevsky, Sutskever and Hinton, 2012]。深層学習は脳の計算モデルに基づいており、その深層学習の畳み込みは視覚の受容野の特性を示している [David Marr, 1982]。この畳み込みは SS 理論 [S. Suzuki, 1994] に端を発し、感性は受容野の反応と個人の経験に基づいた意味付けにより意識される感覚と前提される。

本件においても同様に深層学習の畳み込みを受容野とし、個人の経験に基づいた意味付けをニューラルネットワーク [Cochocki and Unbehauen, 1993] として分離、個人の経験に基づいた意味づけに至る前の楽曲の印象を画一な楽曲の印象と前提する。この構成に感性を表すラベルを付与し、深層学習の学習過程より受容野の特徴を求め、画一な楽曲の印象とし、かつ深層学習をもちいた計算機実験により実験の時間の短縮を行う。

2. 楽曲の感性の特徴モデル

2.1 楽曲の特徴量

楽曲は波形で表される。また、楽曲の印象は音色に起因すると前提する [Hevner, 1937]。ゆえに楽曲の特徴量をスペクトログラムとする。ただし、楽曲の特徴量を鮮鋭とするために、マルコフクラスタをもちいて端点を得ようスペクトログラムの要素の値を先鋭化する。この際、マルコフクラスタのパラメータ a を 3.0、 e を 2.0 とおく。なお、楽曲の教師信号は 1,090 の楽曲からなり、ラベルには感性を表す 90 種の語彙が付与される。この語彙は賑やか、穏やか、爽やかを始めとする情感を表す語彙からなる^{*1}。本稿ではこれを感性語彙と呼ぶ。

2.2 楽曲のグラフ表現

フーリエ級数は正規直交基底関数の正弦波関数の多項式として表せられる。ゆえにラプラス固有関数 $f(w)$ をもちいて、次数行列 D と隣接行列 A からなるラプラシアン行列 L の固有ベクトルより、フーリエ変換に等しい結果を得られる。

$$f(w) = \langle f, e^{j\omega t} \rangle = \int_R f(t) e^{-j\omega t} dt \quad (1)$$

$$e^{-j\omega t} = \frac{\partial^2}{\partial t^2} e^{-j\omega t} = w^2 e^{-j\omega t} \quad (2)$$

$$w^2 = (2\pi\xi)^2 \quad (3)$$

楽曲の特徴量は、スペクトログラムの要素をグラフの頂点、隣接行列 A で任意の近傍の頂点同士を結合した構成をとる。次数行列 D の辺の重みはこれらスペクトログラムの要素の値の相関を重みとする。すなわち楽曲全てのスペクトログラムの要素の相関行列となる。なお、 λ は固有値、 μ_λ は固有ベクトルを表す。以上より、楽曲の特徴量は、スペクトログラムの要素について、任意の近傍の頂点間の値の振動として表される。

$$L = D - A = U\Lambda U^T \quad (4)$$

$$U = [\mu_{\lambda_0}, \dots, \mu_{\lambda_{N-1}}] \quad (5)$$

$$\Lambda = \text{diag}(\mu_0, \dots, \mu_{N-1}) \quad (6)$$

2.3 楽曲の畳み込み

楽曲の特徴量の畳み込みは、[Defferrard, Bresson and Vandergheynst, 2016] らの 128 要素の 2 乗からなるグラフカーネルフィルタ H を 20 次まで展開して畳み込む。グラフカーネルフィルタ H は楽曲の特徴量について逆フーリエ変換 U とフーリエ変換 U^T との積からなる多項式 $U \text{diag}(H(\lambda_H)) U^T f_{in}$ の H であり、次数は振動の大きさを表す。この際に求められる f_{out} が目的とする楽曲の感性の特徴である。

$$P_{\lambda_i} = \sum_{\lambda=\lambda_i} \mu_\lambda \mu_\lambda^T \quad (7)$$

$$P_{\lambda_i} P_{\lambda_j} = \delta(\lambda_i - \lambda_j) P_{\lambda_i} \quad (8)$$

$$f_{out} = H f_{in} \quad (9)$$

$$= \left(\sum_{\lambda_i \in \sigma(L)} H(\lambda_i) P_{\lambda_i} \right) f_{in} \quad (10)$$

$$= U \text{diag}(H(\lambda_H)) U^T f_{in} \quad (11)$$

連絡先: 錠 尚史, 株式会社ソケット, ikari@sockets.co.jp

^{*1} 商用データのために非公開である、必要の際は連絡先まで連絡されたし

2.4 楽曲の感性の個人差を表すニューラルネットワーク

フーリエ級数は級数が収束する前提の元、微分可能である。ゆえにグラフカーネルフィルタは、ラベルと楽曲の特徴量より誤差逆伝搬法をはじめとする学習過程で求められる。この際、グラフカーネルフィルタの荷重は、90種の感性語彙が付与されたラベルと、グラフ表現された楽曲の特徴量を教師信号とし、グラフカーネルフィルタを入力層、1層の中間層を持つニューラルネットワークを出力層とした構成より誤差逆伝搬法をもちいて求められる。上記の f_{out} と1層の中間層を持つニューラルネットワークにおける荷重とは異なり、この荷重は個人の感性の差を表す役割を持つ。

3. 楽曲の感性の特徴

3.1 楽曲の識別の精度

楽曲の感性の特徴モデルをもちいて、1,090の楽曲、ならびに90種の感性語彙を付与したデータセットについて、深層学習の学習過程よりグラフカーネルフィルタを求め、さらに楽曲の特徴量とグラフカーネルフィルタより、全ての感性語彙を満たす楽曲の感性の特徴を求めた。その楽曲の特徴量の識別の精度は96%に達した。この精度について学習回数の変動を加えた[表1]を以下に記す。この際、汎化性能については1,090の楽曲の内の10%をアトランダムに訓練データとした。

表 1: 楽曲の識別の精度

Epochs	訓練正答率	汎化正答率
1	0.0073	0.1431
40	0.8284	0.8110
80	0.9532	0.9394
100	0.9633	0.9514

3.2 楽曲の感性の特徴の俯瞰

得られた楽曲の感性の特徴について、全体の俯瞰を得るために、全ての楽曲の感性の特徴をグラフカーネルフィルタごとに重ね合わせた[図1]を作成した。グラフカーネルフィルタはフーリエ級数展開の多項式に等しく、4次項までで楽曲の感性の特徴がほぼ表されている。

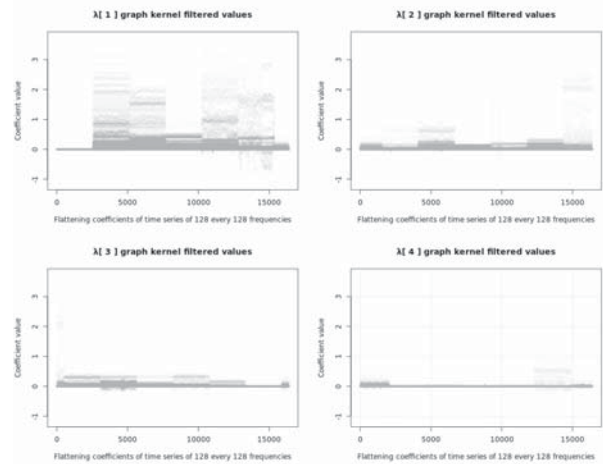


図 1: 楽曲の感性の特徴 (全体)

[図1]の横軸は楽曲の特徴量とグラフカーネルフィルタの積

をとり、128要素の2乗を128要素ごとに1次元に平坦化した軸である。これは左より高周波成分、右に進むにつれて低周波成分であることを示す。縦軸は楽曲の感性の特徴、すなわちグラフカーネルフィルタの係数を表す。この係数は値が高いほど振動が多い。楽曲の特徴量が与えられた際、受容野相当のグラフカーネルフィルタから得られる係数、すなわち楽曲の感性の特徴は、横軸の階数を6とした際に高周波帯、低周波帯、中周波帯のそれぞれ2種の反応よりなる。これは聴覚の基底膜の反応と同じ傾向を示している [Pickles, 1995]。

3.3 特徴の次元圧縮による俯瞰

楽曲の感性の特徴は、グラフカーネルフィルタの係数であるため、グラフカーネルフィルタの128要素の2乗、それを20次持つ。すなわち327,680要素で表される。ゆえに楽曲の感性の特徴を固有値分解し、固有ベクトルより傾向を得る。得られた固有ベクトルは、およそ3種の方向に集約される[図2]。



図 2: 特徴の次元圧縮による俯瞰

3.4 感性語彙と楽曲の感性の特徴

感性語彙は、活性と非活性を表す軸、快と不快を表す軸の直交上で表される [Russell, 1980]。ゆえにそれらの軸に当該する語彙を活性について賑やか、非活性について穏やか、快について爽やかと前提する。不快は不快に相当する感性語彙がないために割愛した。ここで感性語彙に紐づく楽曲の例を挙げる、賑やかは真夏の Sounds good! [AKB48] 他、穏やかは Too Much Rain [Paul McCartney] 他、爽やかは Love Is Stronger Than Pride [Sade] 他である。これらの楽曲の感性の特徴は、賑やかでは、1次と2次に注目すれば高周波帯と中周波帯に高い値が分布し[図3]、穏やかでは、1次に注目すれば低周波帯に高い値が分布[図4]、爽やかでは、1次に注目すれば高周波帯に高い値が分布する[図5]。

ここで感性語彙の直交軸から離れた感性語彙について、楽曲の感性の特徴を求める。非活性と快の中間に位置付けられる感性語彙は、愛しいと前提する。楽曲の例は If You Leave Me Now [Chicago] となる。得られた楽曲の感性の特徴は、1次と2次に注目すれば低周波帯と高周波帯に高い値が分布する[図6]。感性語彙の軸の非活性、すなわち穏やかには低周波帯に高い値が分布し、快、すなわち爽やかには高周波帯に高い値が分布している。ゆえに愛しいは感性語彙の直交軸の和からなると考えられる。また、これらの感性語彙の直交軸の周波帯の傾向は、[3.3 特徴の次元圧縮による俯瞰]の3種の方向を示していると考えられる。

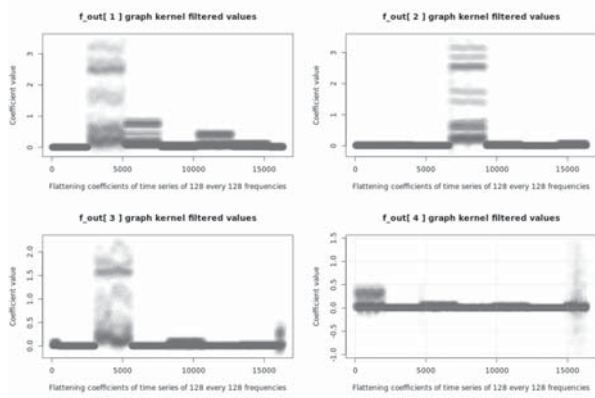


図 3: 楽曲の感性の特徴（賑やか）

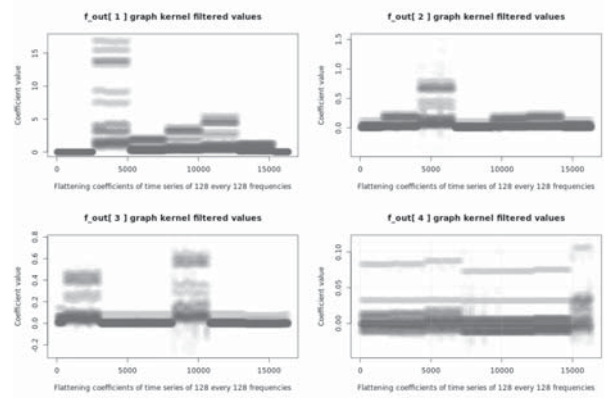


図 5: 楽曲の感性の特徴（爽やか）

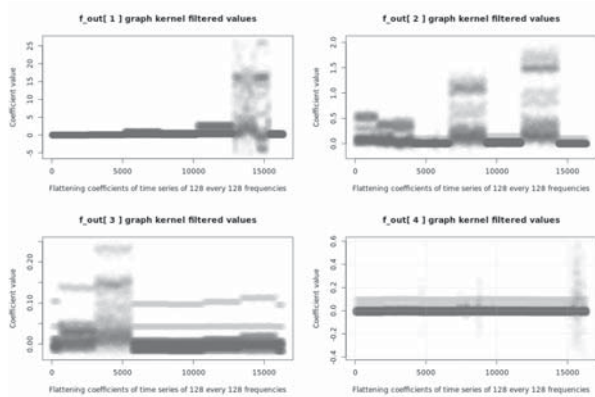


図 4: 楽曲の感性の特徴（穏やか）

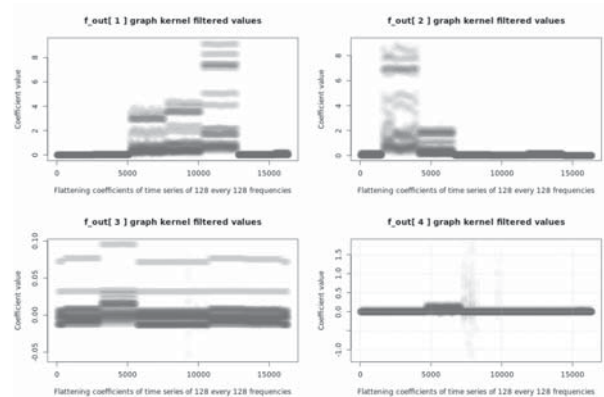


図 6: 楽曲の感性の特徴（愛しい）

3.5 楽曲の感性の特徴と感性語彙の類似性

楽曲の感性の特徴は感性語彙に紐づく。また、楽曲の感性の特徴は数列として表される。ゆえに楽曲の感性の特徴より幾何表現が可能である。以下に恋愛に関する感性語彙が付与された楽曲の集まりの例を記す [図 7]。恋愛を表す感性語彙を持つ楽曲の感性の特徴は、類似した数列となり、近い位置関係が求められている。また、同様に勇気づけに関する感性語彙が付与された楽曲の感性の特徴も類似した数列となり [図 8]、楽曲の感性の特徴は、感性語彙に基づいた類似性を得ていると言える。

会いたくて泣きそう (大好き1.txt.dit, 386)
会いたくて泣きそう (愛1.txt.dit, 386)
失恋、あきらめきれない (切ない3.txt.dit, 396)
失恋、あきらめきれない (愛しい3.txt.dit, 396)
失恋、あきらめきれない (泣ける3.txt.dit, 396)

図 7: 恋愛を表す感性語彙を持つ楽曲の位置関係

集中したい時のワールド (アンセム8.txt.dit, 306)
集中したい時のワールド (ライトな8.txt.dit, 306)
集中したい時のワールド (スリッパ8.txt.dit, 306)
集中したい時のワールド (先進的な8.txt.dit, 306)
集中したい時のワールド (シックス8.txt.dit, 306)
集中したい時のワールド (ヨーロッパい8.txt.dit, 306)
集中したい時のワールド (リード8.txt.dit, 306)
通勤・通学、1日ががんばるぞ、のうた (快活な1.txt.dit, 590)
通勤・通学、1日ががんばるぞ、のうた (アクリル1.txt.dit, 590)
通勤・通学、1日ががんばるぞ、のうた (まっすい1.txt.dit, 590)
通勤・通学、1日ががんばるぞ、のうた (音楽感1.txt.dit, 590)
集中したい時のワールド (キンキン8.txt.dit, 306)
集中したい時のワールド (国際的8.txt.dit, 306)
集中したい時のワールド (小粋8.txt.dit, 306)

図 8: 勇気づけを表す感性語彙を持つ楽曲の位置関係

3.6 楽曲の感性の特徴同士の類似性

楽曲の感性の特徴は、深層学習の学習過程をもちいて求められる。ゆえに楽曲の特徴量より感性語彙の識別が可能である。識別は深層学習に基づくため、識別の結果、得られる楽曲の特徴量の分類の値は、その分類の荷重を表す。本項では、この分類の荷重の値と楽曲の感性の特徴の類似性より、求められる類似楽曲の例を示す。ここでは感性語彙より特徴が顕著な賑やかな真夏の Sounds good![AKB48]、穏やかな Too Much Rain[Paul McCartney]、爽やかな Love Is Stronger Than Pride[Sade]、愛しいの Lady Marmalade[Christina Aguilera] を例にもちいる。それぞれ、賑やかでは summer tail[KEYTALK]、穏やかでは Heaven On The Ground[Jos James]、爽やかでは The Lady Wants To Know[Michael Franks]、愛しいでは Want U Back[Cher Lloyd] が求められた。これら類似楽曲の特徴を、賑やかな類似を [図 9] に、穏やかな類似を [図 10] に、爽やかな類似を [図 11] に記す。

4. まとめ

楽曲の特徴量より、グラフカーネルフィルタと深層学習の学習過程をもちいて、楽曲の感性の特徴を求めた。この楽曲の感性の特徴に基づく識別の精度は 96% に達した。得られた楽曲の感性の特徴を固有値分解した際、その特徴は、おおよそ 3 種の方向に集約される固有ベクトルからなり、おおよそその楽曲の感性の特徴の傾向が示された。また、楽曲の感性の特徴は、

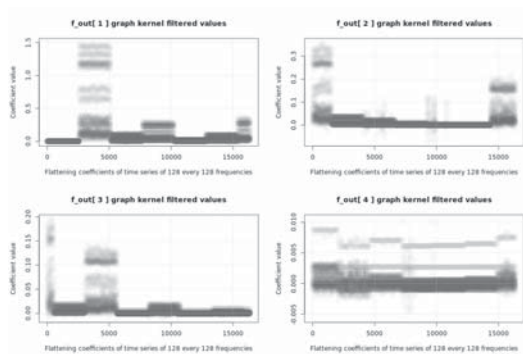


図 9: 楽曲の感性の特徴（賑やかなの類似）

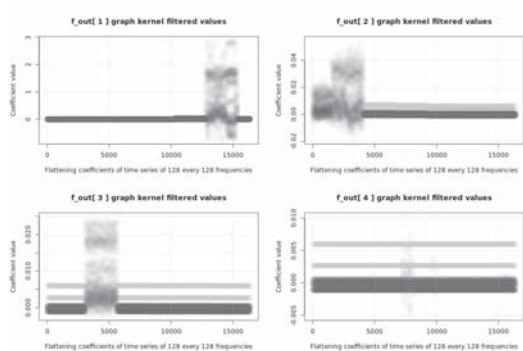


図 10: 楽曲の感性の特徴（穏やかなの類似）

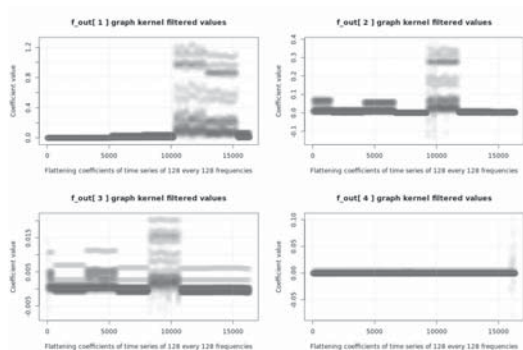


図 11: 楽曲の感性の特徴（爽やかなの類似）

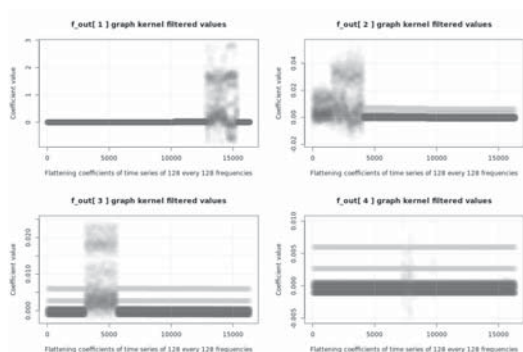


図 12: 楽曲の感性の特徴（愛しいの類似）

賑やかな感性語彙が与えられた際には高周波成分と中周波成分が高く、穏やかな感性語彙が与えられた際には低周波成分が高く、爽やかな感性語彙が与えられた際には高周波成分が高く、感性語彙とそれに紐づく楽曲の感性の特徴の関係が認められ、3 種の方向に集約される固有ベクトルとの関連が示唆された。加えて、感性語彙と楽曲の感性の特徴の関係、楽曲の感性の特徴同士においてそれぞれ類似性が示された。上記より、得られた楽曲の感性の特徴を充たと結論する。

ただし、今回もちいた楽曲の数は 1,090 であり少数であることをふまえ、今後はより大規模な楽曲の感性の特徴を求め、より微細な特徴の獲得を可能とすること、楽曲の感性の特徴と感性語彙に関係が見られることより、その関係を正準相関と仮定し、深層学習の学習過程をもちいた、楽曲の感性の特徴のさらなる追求を今後の課題とする。

5. 謝辞

本件を進めるにあたり、日常の議論を通じての多くの知識や示唆を頂いた株式会社ソケットの皆様には感謝します。協力していただいた皆様へ心から感謝の気持ちと御礼を申し上げたく、謝辞にかえさせていただきます。

参考文献

- [1] Hevner, K. (1935b). Expression in music: A discussion of experimental studies and theories. *Psychological Review*, 42, 186-204.
- [2] C.J.Clopper, E.S.Pearson, (1934). The use of confidence- or fiducial limits illustrated in the case of the binomial, *Biometrika* 26 404-416.
- [3] Krizhevsky, A., Sutskever, I. and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems* 25 1090-1098.
- [4] Marr, D. And Hildreth, E. (1980). Theory of edge detection. *Proc. R. Soc. Long. B*, 207:187-217.
- [5] 鈴木昇一. (1994). パターンのエントロピーモデル, 電子情報通信学会論文誌 (D-II), vol.J77-D-II, no.10, pp.2220-2238.
- [6] A.Cochocki, Rolf Unbehauen. (1993). *Neural Networks for Optimization and Signal Processing* 1st. John Wiley and Sons, Inc.
- [7] Hevner, K. (1937). The affective value of pitch and tempo in music. *American Journal of Psychology*, 49, 621-630.
- [8] Michal Defferrard, Xavier Bresson, Pierre Vandergheynst. (2016). Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Advances in Neural Information Processing Systems* 29.
- [9] J.O.Pickles, 谷口郁雄監訳. (1995). 聴覚生理学, 二瓶社.
- [10] J.A.Russell. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, Vol. 36, pp. 1161-1178.