視覚的印象を表現する楽曲特徴量の抽出および楽曲生成手法

陳 昀劭† 清木 康†

†慶應義塾大学環境情報学部 〒252-0882 神奈川県藤沢市遠藤 5322

E-mail: †{t16577yc, kiyoki}@sfc.keio.ac.jp

あらまし 「見える音とは何か」の探求においては、人間が新たな情報を受け取る際に、自らの基準による感性評価が重要である。たとえ異なる形式を持つ情報であっても、それらに共通している印象に基づいた記憶想起は可能である。そうした異種メディアを対象とした複数の処理能力の連携性に着目し、本稿では動画における視覚的印象の時系列を分析し、相応した印象を有する音楽的要素の連想検索および抽出方式の提案を行う。さらに、その出力結果に基づいて楽曲を生成するクロスメディアシステムの実現方式を示す。

キーワード 感性情報処理、楽曲生成、マルチメディア

1. はじめに

共感覚とは、一つの刺激に複数の感覚が引き起こされる現象である。[1]共感覚を有する人は共感覚者と呼ばれ、音や文字などに色が見える。共感覚者以外の人も、ある情報を自分なりに評価・解釈する際に、相似したプロセスが行われる。例えば、ピッチの高い音を聞かされた後、どれに近いかという質問に対し、人は丸まった図形より尖った図形を選択する傾向があるという実験結果が発表されている。[2][3]心理学分野ではこの現象を、複数の感覚間に共通したある心理的性質の存在による結果とし、「通様相性」と呼んでいる。[4]共感覚との相違としては、通様相性は普段の意識にのぼらず、感覚間の関係性が知覚できる程度である。[1]

共感覚なり通様相性なり、共通の心理的性質による、異なる形式で記憶した情報への関連付けは、人間の脳内における異種メディアを対象とした情報処理に連携性があることを示唆している。こうした連想のプロセスと同様に機能するシステムを利用すれば、ある情報から受け取る感性を別のメディアで再現することが可能となる。感性に基づくメディア間の変換に関する研究として、画像や単語などの印象をもとに音楽を検索、もしくはアレンジするシステムが発表されている。[5][6]その中には、事前に準備したプレーリストの中から曲を推薦する方式が多く存在する。[7][8][9]しかし、音楽における感性は常に同一ではなく、曲の展開とともに次々と変化していくのである。そのため、1 曲単位で検索・推薦を行うと感性の時間的変化を確実に反映することが困難である。そこで本稿では、楽曲そのものではなく、音楽を構成する要素のデータ、すなわち楽曲特徴量を検索対象とした方式および、個別に抽出された特徴量を利用して楽曲を生成する手法を提案する。また、音楽と同様に展開が作品のクオリティーに直結するメディアである点から、変換対象として動画がふさわしいと考えられる。動画における感性の時系列を基準に、その印象の変化を具現化できる楽曲を生成する。

2. システム構成

本システムは、動画用感性分析ユニット、感性類似度解析ユニット、楽曲生成ユニットによって構成される。意味的連想検索方式[10][11]と時系列データ処理の技術を用い、動画から受け取る感性をもとに楽曲を生成する。

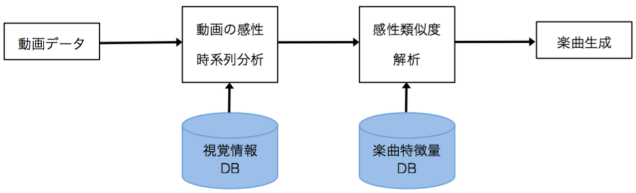

図 1 システム構成

3. 視覚情報の感性評価機構

最初に、動画用感性分析ユニットで実行される感性時系列分析の方式を示す。

3.1 動画における視覚情報の分析

動画における視覚的印象の変化を計算可能にするため、視覚情報の感性評価を行う。本システムは、視覚情報として色彩と画面に含まれている物体を感性評価の対象とする。

色彩については、一定間隔で動画のフレームを取得し、フレームごとに k 平均法（k-means）にてクラスタリングすることで、画面から主な色の組み合わせを特定する。（以下、配色データ）

物体は、Google 社が開発した機械学習ライブラリである TensorFlow[12]を利用して検出する。(以下、物体データ) なお、フレーム取得の時間間隔は、配色の分析と物体検出それぞれ設定可能である。

3.2 感性成分の生成手法

通様相性の特徴からわかるように、同じ印象語群で視覚情報と音声を評価すれば人間の記憶想起から発想を得た感性評価のメカニズムをシステム上で実現可能である。

色がもたらす印象を計量するために、小林が発表したカラーイメージスケール[13]を参考にして soft-hard 軸と warm-cool 軸からなる二次元空間を構築した。配色データをカラーイメージスケールの空間に写像し、配色の印象ベクトルが得られる。

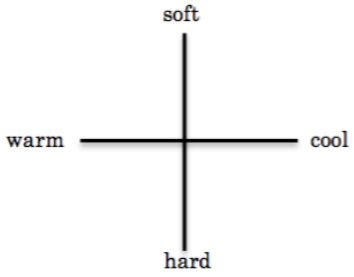

図 2　カラーイメージスケールのベクトル空間

音楽を構成する各要素が与える印象を言語化した実例として、Kirnberger の著述があげられる。[14][15] Kirnberger が音程の表現に用いた印象語群にてあらかじめ各物体を評価することで、物体データから印象ベクトルを取得する。

上述の手法で配色および物体を評価対象とした感性時系列を生成する。

4. システムの実現方式

動画から抽出された感性時系列を基準とした、同じ感性を共有する音楽的要素の検索は、感性類似度解析ユニットによって実現する。本ユニットは楽曲特徴量データベースおよび感性類似度比較機能により構成される。

4.1 楽曲特徴量データベースの構築

前述のように、音楽における感性がダイナミックな点に焦点を当て、本研究においては曲単位ではなく、音楽を要素別に細分して楽曲特徴量データベースを構築した。楽曲特徴量データベースには音楽的要素と個々の感性成分が格納されている。曲中の時間的変化に重点を置いているため、音楽的要素として調および音程を感性の比較対象とする。

カラーイメージスケールは印象の距離を表す空間にて色を評価したモデルであり、この空間では色が与える印象は数値化される。Charpentier の調性格に関する論述[15][16]をもとに各調性を同じ空間に写像することで調の感性成分が得られる。

音程がもたらす印象については Kirnberger が用いた印象語群を利用して感性成分を記述する。複数の印象語を含んでいる音程が多く存在するため、二次元空間ではなく各印象語の強度をもとに 4 段階評価を行い (以下、印象強度)、音程の感性成分を多次元空間にてベクトル化する。

表 1　印象強度

印象	評価
強く持っている	3
持っている	2
ややある	1
なし	0

表 2　楽曲特徴量データベースの一部

Interval	Semitones	anguished	sorrowful	agreeable	pathetic	languishing	painful	rejoiced	plaintive	joyous
Perfect_unison[ascending]	0	0	0	0	0	0	0	0	0	2
Augmented_unison[ascending]	1	2	0	0	0	0	0	0	0	0
Minor_second[ascending]	1	0	2	0	0	0	0	0	0	0
Major_second[ascending]	2	0	0	2	2	0	0	0	0	0
Augmented_second[ascending]	3	0	0	0	0	2	0	0	0	0
Diminished_third[ascending]	2	0	0	0	0	1	0	0	0	0
Minor_third[ascending]	3	0	2	0	0	0	2	0	0	0
Major_third[ascending]	4	0	0	0	0	0	0	2	0	0
Diminished_fourth[ascending]	4	0	0	0	0	0	2	0	2	0
Perfect_fourth[ascending]	5	0	0	0	0	0	0	0	0	2

4.2 動画と楽曲特徴量の感性類似度計量系

同じ印象空間を基準にすることによって、動画と音楽それぞれ含む感性成分は比較可能になる。

動画における配色は色相、同時に映る色の数、そして個々の比率などによって緻密に画面全体の印象に影響を及ぼす。一方、音楽においては音階 (スケール) や和音など、様々な手がかりから調性を推定できる。このように、感性への影響が全面的である点が共通しているため、配色と調の感性を比較する。

それに対して画面にある物体はすべて均一に感性を誘発するのではなく、一部の物体に全体の印象が左右される。人間は静止の状態を続ける物体を背景の一部と認識し、背景が動体に付随した際のみ感性の経時変化が取得可能である。一つの動画に動体が多数存在する可能性が高く、印象語が頻繁に変動することが予想されるため、メロディーラインや和声進行のような連続的に変化する要素がふさわしい。こうした要素には音の前後関係、すなわち音程の上下が決定的影響を与える。よって音程を物体がひき起こす感性の比較対象とする。

画面の配色と各調性、両者の印象ベクトルのコサイン類似度を計算し、コサイン類似度のもっとも高いものは色彩の印象を表現するに最適な調として出力される。コサイン類似度を用いる理由としては、カラーイメージスケールの空間において角度が印象語の関連性をより適切に表現できるためである。

検出された物体と各音程との感性類似度比較機能では、多次元空間における印象ベクトルの計算を行うので、内積、コサイン類似度、ユークリッド距離、三つの相関量から選択可能である。評価基準となる印象語が n 個存在し、物体の印象ベクトルが x, 各音程の印象ベクトルが y_m の場合、以下の関係が満たされる。

$$x = (x_1,\ x_2,\ \ldots,\ x_n)$$

$$y_m = (y_{m1},\ y_{m2},\ \cdots,\ y_{mn})$$

$$\text{内積}\ <x,\ y_m> = \sum_{i=1}^{n} x_i\, y_{mi}$$

$$\text{コサイン類似度} = \frac{<x, y_m>}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_{mi}^2}}$$

$$\text{ユークリッド距離} = \sqrt{\sum_{i=1}^{n}(x_i - y_{mi})^2}$$

	Perfect_unison[ascending]	Augmented_unison[ascending]	Minor_second[ascending]	Major_second[ascending]
person	0.240192	0.138675	0.277350	0.498290
bicycle	0.547723	0.000000	0.000000	0.223607
car	0.444444	0.000000	0.000000	0.272166
motorcycle	0.485071	0.000000	0.000000	0.297044
airplane	0.425628	0.147442	0.000000	0.312772
bus	0.300607	0.534522	0.000000	0.188982
train	0.222222	0.577350	0.192450	0.136083
truck	0.144338	0.250000	0.000000	0.000000
boat	0.258199	0.000000	0.000000	0.316228

図 3　物体と音程とのコサイン類似度（一部）

最大の感性類似度を有する音程は、検出された物体に相応した印象を与えうる要素として抽出される。

```
************************ QUERY:boat ************************
——InnerProduct: 10 ——
['Perfect_octave[ascending]']
——CosineSimilarity: 0.6454972243679028 ——
['Perfect_octave[ascending]']
——EuclideanDistance: 3.4641016151377544 ——
['Perfect_octave[ascending]']
```

図 4　音程の抽出結果（一例）

4.3 感性時系列による楽曲生成

感性類似度解析ユニットによって調と音程の時系列データが出力される。(以下、出力結果) このデータは動画における色彩および物体と同じ感性成分を共有する。以下、出力結果を用いて楽曲を生成する手法を示す。

本システムは、出力結果における音程の時系列に沿った和声進行によって曲を紡ぎだす。具体的にはダイアトニック・コードを利用する。[17]ダイアトニック・コードとは、全音階の音のみで構成され、主調の性格を自然に表現できるコードである。短調の場合には響きを考慮し、III 度の際のみ自然短音階を用い、その他は和声的短音階を基準とする。

表 3　楽曲生成に用いるダイアトニック・コード

長調	短調
$I_{\triangle 7}$	$Im_{\triangle 7}$
II_{m7}	$II_{m7^{(\flat 5)}}$
III_{m7}	$\flat III_{\triangle 7}$
$IV_{\triangle 7}$	IV_{m7}
V_7	V_7
VI_{m7}	$\flat VI_{\triangle 7}$
$VII_{m7^{(\flat 5)}}$	VII_{dim7}

最初の和音は、再生位置 0 分 0 秒の調と同じ時刻の音程によって決定される。例えば、調が C メジャー (ハ長調)、音程が完全 5 度上行の場合、C メジャーの I 度と完全 5 度で隔てられる上方の G_7 が最初の和音となる。それ以降は直前の根音を基準に、出力結果における音程の時系列に従って移動させることで和声進行を生成する。なお、調の時系列は根音以外の構成音に反映される。

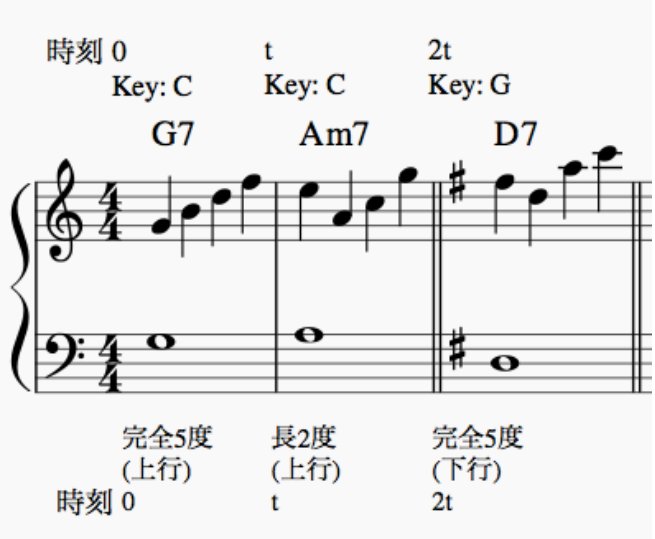

図 5　和声進行の生成手法

出力された和声進行に基づいて低音部に根音を配置し、高音部では分散和音を演奏させることでメロディーラインが形成される。なお、メロディーがなめらかに生成されるように、コード間の遷移に際して直前の分散和音の末尾にある音により近い音を先頭に配置する。

上記の手法で動画がもたらす視覚的印象を体現する楽曲が生成可能となる。動画と同時に再生することが

想定されるため、フレームの取得と同じ時間間隔で出力結果を音声に変換する。楽曲生成ユニットは音響合成用プログラミング言語であるSuperColliderにて実装されている。[18]

5. 考察

今回の実験結果から、感性類似度の高い楽曲特徴量を抽出できることがわかった。実用性を再考した結果、改善点として次の2点があげられる。

まず、感性類似度計量系における同等な相関量に対する処理である。本方式は、画面との感性類似度が最大な楽曲特徴量が抽出されるように設計されているが、実際には等値の相関量をもつ特徴量が存在し、複数個同時に抽出されることがある。この現象は特に、画面にある物体と各音程との感性類似度を比較する際に多発している。その原因は評価基準となる印象語の個数にあると考えられる。本方式において用いられるKirnbergerの印象語は30個あり、この高次元の印象空間に対して印象強度は4段階で評価されている。実際に感じられる強度の差分をより精確に表すには、評価手法を改良する必要がある。また、相関量が等値の場合に行われる処理として、重み付けがあげられる。同等な相関量は、本計量系において感性類似度が等しいことを意味するのであるが、メロディーラインの生成がスムーズに実行されるように、複数の音程が同時に抽出される出力方式を回避すべきである。

次に、楽曲生成に用いる音楽的要素である。抽出された音程の時系列に沿ってダイアトニック・コードを演奏させるのが今回の手法であった。しかしながら、その音程が必ず全音階の音に導くことは保証されていない。その解決策として、経過音を介して全音階に含まれる音へ移行させる処理、または代理和音の導入などの和声的な手法が有効だと予想される。

6. 結論・今後の展望

本稿では、動画における視覚情報を対象とした感性分析および、動画の感性的特徴を音声の形式で再現可能な音楽的要素の抽出方式、そして出力された時系列データを用いて楽曲を生成する実験手法を示した。

今回は、印象語の選択に際して音楽家による著作を参考にした。音楽における各要素の特性が的確に把握されている利点はあるが、一般人の感性とどれほど合致するかについて検討する余地がある。今後の課題として、既存のデータを用いた計量にとどまらず、ユーザーの感性をシステムに反映するパーソナライゼーションがあげられる。

さらに、リズムを筆頭に、他の音楽的要素を導入することで、多階層構造で楽曲を生成できる手法の有効性について検証したい。将来的には、ローカルの動画に限らず、ライブ配信などのストリーミングサービスにも適用可能なマルチメディアシステムを目標とする。

参 考 文 献

[1] 長田典子「音を聴くと色が見える：共感覚のクロスモダリティ」, 日本色彩学会誌, Vol. 34, No.4, pp. 348-353, 2010

[2] W. Köhler, *Gestalt Psychology* , Liveright, New York, 1929

[3] V.S. Ramachandran and E.M. Hubbard, “Synaesthesia—A Window Into Perception, Thought and Language”, *Journal of Consciousness Studies*, Vol. 8, No. 12, pp. 3-34 , 2001

[4] 長田典子, 藤澤隆史「共感覚の脳機能イメージング」, システム／制御／情報, Vol. 53, No.4, pp. 149-154, システム制御情報学会, 2009

[5] 仲村哲明, 内海彰, 坂本真樹「色彩想起と歌詞の関係に基づく楽曲検索」, 人工知能学会論文誌, Vol. 27, No.3, pp. 163-175, 人工知能学会, 2012

[6] 大山喜冴, 伊藤貴之「DIVA：画像の印象に合わせた音楽自動アレンジの一手法の提案」, 芸術科学会論文誌, Vol. 6, No.3, pp. 126-135, 2007

[7] 黒瀬崇弘, 梶川嘉延, 野村康雄「感性情報を用いた楽曲推薦システム」, 第 14 回データ工学ワークショップ（DEWS2003）, 8-P-6, 2003

[8] 北川高嗣, 中西崇文, 清木康「楽曲メディアデータを対象としたメタデータ自動抽出方式の実現とその意味的楽曲検索への適用」, 電子情報通信学会論文誌, Vol. J85-D-I, No.6, pp.512-526, 2002

[9] 桐本篤, 佐々木史織, 清木康「音楽と感性語の相関量計量による環境状況コンテクスト対応型音楽推薦システムの実現」, 第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM2009）, E5-2, 2009

[10] 清木康, 金子昌史, 北川高嗣「意味の数学モデルによる画像データベース探索方式とその学習機構」, 電子情報通信学会論文誌, Vol. J79-D-II, No. 4, pp. 509-519, 1996

[11] 清木康「感性や意味を計量するデータベースシステム」, KEIO SFC JOURNAL, Vol. 13, No.2, 2013

[12] TensorFlow https://www.tensorflow.org/

[13] 小林重順, 日本カラーデザイン研究所, 『カラーイメージスケール』,講談社,2001

[14] J. P. Kirnberger, *Die Kunst des reinen Satzes*, Berlin, 1776, pp. 103-104, reprint : Olms, 1988.

[15] P. A. Clerc, “Discours sur la Rhétorique musicale”, Haute École de Musique de Genève, Genève, 2001

[16] M.A. Charpentier, “Règles de composition par M. Charpentier“ in *Marc-Antoine Charpentier* by Catherine Cessac, Fayard, Paris, 1988
[17] ダイアトニック・コード, 洗足オンラインスクール, 洗足学園音楽大学
https://www.senzoku-online.jp/theory/chord/ch01_01.html
[18] SuperCollider　https://supercollider.github.io/