2C3-OS-20a-2in1

文書や画像の印象にもとづく楽曲生成

Music synthesis based on impression of input documents and movies

伊藤貴之*1
Takayuki Itoh

お茶の水女子大学
Ochanomizu University

Our laboratory had studies on techniques for music synthesis based on impression and input documents and movies. This paper introduces these projects. The paper also discusses remaining issues on these studies and future visions.

1. はじめに

音楽は多くの場面において多くのメディアとあわせて制作される. オペラやミュージカルといった劇場芸術では舞台進行にあわせて音楽が制作される. 映画やテレビでは BGM として音楽が制作される. このようなメディアに音楽を付与する, という行為は一般消費者にも身近なものになった. 例えば最近では, スマートフォンやデジタルカメラで撮影した動画に, BGM をつけて SNS などで共有する, といった形で音楽を用いるサービスも知られている.

しかし自分の感性に合った音楽を付与するという行為は必ずしも手軽なものではない. 動画に BGM をつける作業をひとつとっても, その動画に合うと感じる音楽を探す作業は必ずしも容易ではないし, それが全く見つからない状況も想像しうる. また特に, SNS などで音楽付きメディアを公開する際には, 他者が聞いたことのない音楽を付与したい, 著作権上の問題が発生しない音楽を付与したい, と考える機会もあるだろう. しかし, そのような音楽を自分で制作するには一定のスキルが必要であり敷居が高い. さらに, 自分の嗜好に合わせて付与した音楽が鑑賞者にも共感してもらえるとは限らない. この問題を解決する究極的な手段として, 各々の鑑賞者に自動音楽付与機能を搭載させることが考えられる.

著者の研究室では, 文書や動画を与えられて, それに合うと感じる音楽を自動制作する研究[Kanno2015][清水 2016]に取り組んできた. この研究ではサンプルとなる音楽素材をいくつか用意した上で, 入力メディアとなる文書や動画の印象を推定し, それと印象が近いと推定される音楽素材をマッシュアップすることにより, その場で音楽を合成する. この研究では入力メディアや音楽素材に対する個人の印象を学習し, 個人の嗜好に合わせて音楽素材を自動選択する, という立場にこだわってきた. 本報告ではこれらの研究について紹介するとともに, これらの研究の音楽自動生成分野における立場と問題点, そして今後の展望について議論する.

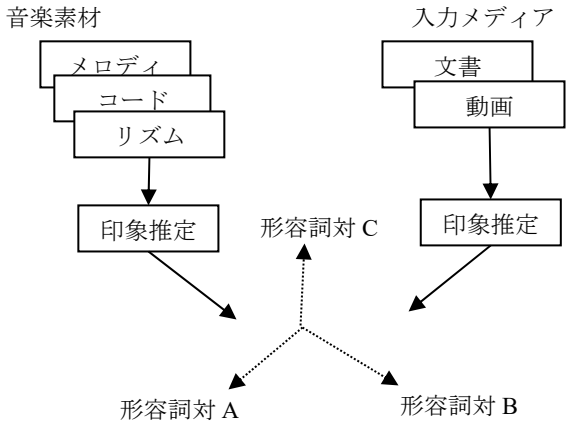
2. 本研究の共通フレームワーク

図 1 に本研究の提案手法に共通する処理手順を示す.

提案手法では学習段階として, サンプル音楽素材(メロディ, コード進行, リズムパターン)をユーザに聴かせ, その印象を数値で回答させる. ここで提案手法ではあらかじめ, SD 法にもとづいて数種類の形容詞対を用意し, 各形容詞対についてその適合度を数値で回答させる. 例えば「暗い⇔明るい」という形容詞対があったときに, 例えば-1.0 と回答したら最も暗い, +1.0 と回答したら最も明るい, というように数値範囲を定義する. この数値を本報告では「印象値」と呼ぶ. 以上の定義にもとづいたユーザ回答結果から, 音楽素材の特徴量と印象値の関係を学習させる. この結果として, 新しく与えられた音楽素材について, 特徴量から印象値を推定する仕組みが成立する. この仕組みに従って, ユーザに聴かせたサンプル音楽素材以外の音楽素材についても, 各々の印象値を推定する. なお, 音楽素材の特徴量から印象値を推定するための学習処理として, 著者らはまず単純な線形重回帰分析を適用したが, のちに自己組織化マップ(Self-Organizing Map: SOM) を用いた学習処理に差し替えている.

また提案手法ではこれに加えて, 入力メディア(文書または動画)の印象値を推定する仕組みを用意する. この仕組みについては次章で後述する.

以上の学習段階が終了したら, 任意の入力メディアに対して音楽素材を自動選出する. まず任意の入力メディアの印象値を推定し, 続いて音楽素材の中でその印象値に最も距離が近いものを選出する. 選出された音楽素材を合成することで, 任意の入力メディアに印象が近いと思われる楽曲を生成する.



各形容詞対での印象値を軸に割り当てた多次元空間で入力メディアに距離の近い音楽素材を選択する
→音楽素材を合成して楽曲を生成する

図 1 本研究に共通する処理手順

連絡先:伊藤貴之, お茶の水女子大学理学部情報科学科, 東京都文京区大塚 2-1-1, itot@is.ocha.ac.jp

なお提案手法において，音楽素材はメロディ・コード進行・リズムパターンともに 1 チャンネルのみの MIDI ファイルで提供されているものとする．各々の特徴量は MIDI ファイル中のノートオン・ノートオフイベントから算出可能な変数を採用している．音楽素材の合成では単純に，新しく生成される MIDI ファイルの各チャンネルにメロディ・コード進行・リズムパターンを複製している．

3. 各手法に固有の処理手順

著者の研究室では，前章で述べたフレームワークに沿って，入力文書からの音楽生成[Kanno2015]，入力動画からの音楽生成[清水 2016]に取り組んできた．本章ではその各々に固有の処理手順について述べる．

3.1 入力文書からの音楽生成

著者の研究室では，入力文書の印象を推定し，これに印象が近いと思われる音楽を自動生成する，という研究に着手した．この研究では，入力文書の印象推定に感性極性抽出手法[高村 2006]を用いている．著者らは前章で述べた形容詞対の各々を種単語として，あらかじめ辞書に登録された各単語と種単語の距離を算出し，これを各単語の感性極性値とし辞書に付記する．以上の処理を事前処理として各々の形容詞対について実施する．以下，感性極性値が付記された辞書を「感性極性辞書」と称する．

入力文書が与えられると，本手法では形態素解析によって文書を単語に分割し，そのうち名詞・動詞・形容詞・副詞を抽出する．抽出された単語について感性極性値を参照し，その重み付き平均をとることで入力文書の印象を推定する．単語の重みについては tf/idf などいくつかの手法を試してみたが，現在も検討中である．

3.2 入力動画からの音楽生成

著者の研究室では，入力動画の印象を推定し，これに印象が近いと思われる音楽を自動生成する，という研究に着手した．この研究では入力動画の印象推定においても，音楽素材の印象推定と同様に，ユーザにサンプルを提示してその特徴量と印象値の関係を学習する，というアプローチを採用した．

著者らの以前の実験結果[Ohyama2007]にて，画像に対して閲覧者が受ける印象は，色などの低水準な特徴量から影響を受ける人と，被写体や背景などが有する高水準な知識から影響を受ける人にわかれる，ということが示唆されている．このことから著者らは，以下の 2 種類の印象推定方法により印象値を算出し，その重み付き和により入力動画の印象値を推定している．

低水準特徴量からの印象推定

事前実験として被験者にいくつかの動画を閲覧させて，いくつかの形容詞対を提示して印象値を回答させ，何らかの印象値と相関が見られる動画特徴量を採用し，何らかの動画特徴量と相関が見られる形容詞対を採用する．続いて各ユーザにサンプル動画を提示して，採用された形容詞対に対して印象値を回答させ，採用された動画特徴量との関係を学習する．

高水準情報からの印象推定

まず入力動画の被写体を推定する．著者らの実装では手入力で被写体をタグ付けしていたが，一般物体認識手法を適用することである程度の自動化が可能である．続いて，あらかじめ採用されたいくつかの形容詞対について，その被写体名との単語間距離を算出し，算出結果から印象値を算出する．著者らの実装では，形容詞対と被写体名との距離算出に word2vec を用いている．

4. 実験

前章までに紹介した両手法ともに，以下の手順で実験を進めた．

1) 事前調査

先行研究を参考にして特徴量や形容詞対の候補を列挙した．そして多数の被験者を対象に，サンプル音楽素材やサンプル動画を提示し，形容詞対を提示して印象値を回答させた．印象値を集計し，音楽特徴量や動画特徴量との相関を算出した．結果として，何らかの印象値と相関が見られる特徴量，あるいは何らかの特徴量と相関が見られる印象値を残し，それ以外の特徴量や印象値を破棄した．

2) 学習

あらためて提案手法のユーザという意味での被験者を数名の学生に依頼し，サンプル音楽素材を聞かせた．そして 1)で選択した各形容詞対を提示し，印象値を回答させた．入力動画からの音楽生成においてはサンプル動画についても同様に印象値を回答させた．そして回帰分析または SOM を適用することで，特徴量と印象値の関係を学習させた．

3) 評価実験

2)で用いたサンプル音楽素材に加えてさらに多くのサンプル音楽素材を用意し，2)の学習結果に従ってサンプル音楽素材の印象値を推定する．続いて入力メディア（文書または動画）を与えてその印象値を推定し，印象値が近い音楽素材を選んで合成する．この結果として生成された楽曲を被験者に聞かせ，入力メディアと印象が合致するかを評価させ，さらにその理由についてフリーコメントを書かせた．

実験結果の詳細については各手法の筆頭著者の修士論文に掲載されている．いずれの手法も，特定の入力メディア，特定のサンプル楽曲において評価が低い傾向が見られた．

5. 議論と課題

本章では本研究が残したいくつかの問題点について議論するとともに，その解決案についても議論する．

5.1 個人の嗜好に合わせた音楽生成に関する問題

本研究は多くの大衆が満足するように音楽を生成するのではなく，「ユーザ個人の嗜好に合わせて音楽を生成したい」という動機から出発した研究である．

既に公開されている映像や文章と音楽の組み合わせからその傾向を倣って音楽を生成したいのであれば，YouTube などの投稿サイトをクロールして得られるメディアの特徴量を深層学習するなどの方法が考えられる．またクラウドソーシングなどの方法によって特定のメディアに対する大衆的嗜好を最適化する研究は最近になって活発化している[Koyama2014]．これらの研究と違って本研究は，入力メディアと音楽との組み合わせに関する個人の嗜好を学習する工程から出発しないといけない．研究開始時から多少は想像がついていたことであるが，この条件設定は非常に難易度の高い条件設定であった．

本研究の問題設定は，あるユーザが当該システムを使用し始めた瞬間には教師信号が全く存在しない，という意味で典型的なコールドスタート問題である．例えば「個人ユーザの音楽鑑賞の嗜好を学習する」というような日常性の高い問題であれば，ライフログ分析などの手法によって日常生活の中で個人ユーザの嗜好を学習することも多少は可能である[Uno2014]．しかし本研究の「入力メディアに音楽を付与する」という工程にそこまでの日常性はない．結果として，システムが個人ユーザの嗜好を学習するために一定のタスクを与える必要があると考えた．しかし

そのタスクに過剰な労働量を課するようでは実用性が損なわれるため，小さいタスクから集まった少量の教師信号からどの程度の満足度を実現できるか，という実験が本研究のキーポイントとなった．

ここまで本研究では特徴量と印象値の関係を学習するために，線形重回帰分析および SOM を用いているが，特徴量と印象値の関係を線形近似できるとは限らない．かといって非線形回帰分析を適用できるほど多量の教師信号をつくるようなタスクは重すぎる．このトレードオフを解決するために，少量の教師信号を前提とした最近の学習手法を適用するなどが必要になると考えられる．

なお「ユーザの嗜好を学習するために一定のタスクを必要とする問題」は本研究以外にも多数ある．画像の自動加工はその典型例である．画像自動加工の研究では近年，学習のためのユーザタスクを軽減するための工夫が議論されている．一例として小山らが提案した SelPh [Koyama2016]では，画像特徴量と画像加工パラメータとの関係に関する学習結果の自信度をユーザに提示することで，学習の達成度を示し，ひいてはタスクの適正化を実現しようとしている．また斉藤らが提案した CrowdRetouch [斉藤 2016]では，初期ユーザによる画像加工タスク結果からあらかじめ初期ユーザをクラスタリングしておき，新規ユーザは学習タスクにおいてどの初期ユーザクラスタに属するかのみを判別することで，新規ユーザの嗜好を学習するためのタスクを軽減している．画像自動加工の研究におけるこれらのアイディアは，音楽自動生成に関する本研究においても参考にもなるものと思われる．

5.2 印象値回答のための形容詞対の選択の問題

本研究では入力メディアと音楽素材の印象値を推定し，印象値間の距離を算出するというアプローチをとってきた．一方で，音楽素材の特徴量空間を入力メディアの特徴量空間に写像し，印象値を介在せずに両者の距離を算出するというアプローチも考えられるはずであったが，本研究では採用しなかった．理由の一つとしてタスクの増加回避があげられる．例えば n 個サンプル入力メディアと m 個のサンプル音楽素材があったときに，各々の印象を回答させるタスクは $O(n+m)$となる．それに対して，サンプル入力メディアとサンプル音楽素材の相性を直接回答させるタスクは $O(nm)$となる．この観点から本研究では，両者の相性を直接回答させるタスクの採用を避けることにした．

印象値を回答させるというタスクは心理学をはじめとして多くの分野において活用されているタスクであるが，しかし本研究において必ずしも的確に回答されていたとは限らない．大きな要因として，形容詞対への解釈に個人差があるという問題がある．例として「明るい⇔暗い」という形容詞対を音楽素材に適用すると，ある被験者はコード進行が長和音か短和音かによって明るさを解釈し，またある被験者は音程が高い・音色がひきたつ，といった別の要因によって明るさを解釈する．このように形容詞対の解釈に個人差があると，それに対応する特徴量にも差が生じる．現在の我々の実装では 4 章の 1)で述べた通り，多数の被験者回答の集計結果にもとづいて形容詞対と特徴量を選択しているが，本来ならこれも個人ごとに選択しないといけないのかもしれない．

5.3 音楽特徴量の選択

本研究では MIDI 形式ファイルの合成によって楽曲を生成しており，MIDI 信号から読み取り可能な情報をもとにして音楽特徴量を算出している．言い換えれば，この特徴量は楽譜情報生成時点（作編曲時点）での特徴量であり，録音前の特徴量であるとも考えられる．

本研究の実験時に収集したフリーコメントを考察すると，この音楽特徴量が裏目に出ているケースが散見される．具体的にいうと，メロディの音列から類推されるストーリー性・コード進行やリズムパターンから類推される音楽ジャンルなどに全く言及せず，シンセサイザーから発音される音色に関してのみ言及している例が見られる．この結果から，音楽鑑賞時に，録音してはじめて確定する音響特徴のみから音楽の印象を感じ取る人がいることが示唆される．このようなユーザに対して MIDI 信号から算出可能な音楽特徴量を適用するのは効果的ではない．音響特徴量から音楽素材の印象を推定する方法についても議論が必要であると考えられる．

5.4 単語の数値化

本研究ではサンプル音楽素材およびサンプル入力メディアから受ける印象を個人ユーザごとに学習するタスクを設けているが，このタスクを介さずに印象を推定している工程も存在する．現時点での本研究の実装では以下の 2 点

- 入力文書の印象推定のために各単語の感性極性を辞書化する処理
- 入力動画に写る被写体の印象推定のために被写体名と形容詞対の距離を算出する処理

において，ユーザごとのタスクを介さずに印象を推定している．しかし本研究での実験結果でのフリーコメントを考察すると，この点にも問題が多くみられる．入力文書の文脈や入力動画のストーリーを考慮せずに，単語が与える印象だけから入力メディアの印象を類推するのは限界がある，というのが現時点での結論であり，この点での改良も議論する必要がある．

6. まとめ

本報告では，文書や動画といった入力メディアに印象の近い楽曲を生成するという研究課題について，著者の研究室の取り組みを紹介し，その問題点について論じた．現時点での本研究は完成度の低い状況であり問題点も山積しているが，類似研究に取り組む方々にとって本報告が何らかの参考になることを望みたい．

ここ数年のビッグデータおよび機械学習の発展により，音楽自動生成の研究においては特に，集合知や実績にもとづいて多くの聴衆が納得する楽曲が生成できるようになったと感じる．一方でパーソナライズされた楽曲生成という課題はそれと独立に近い形で学術的価値を有するものと考える．個人タスクというスモールデータにもとづいたパーソナルな楽曲生成の研究が今後ますます活性化されることを願っている．

謝辞

非常に難易度の高い研究課題を自ら発案し 3 年間果敢に遂行した当研究室修了生（菅野沙也，清水柚里奈）に敬意を表する．また本研究の共同研究者として暖かい支援をいただいた，東京工業大学高村大也先生，明治大学嵯峨山茂樹先生，シドニー大学高塚正浩先生に感謝の意を表する．

参考文献

[Kanno2015] S. Kanno, T. Itoh, H. Takamura: Music Synthesis based on Impression and Emotion of Input Narratives, Sound and Music Computing Conference (SMC2015), pp. 55-60, 2015.

[清水 2016]　清水, 菅野, 伊藤, 嵯峨山, 高塚: 動画特徴量からの印象推定に基づく動画 BGM の自動素材選出, NICOGRAPH 2016, C-6, 2016.
[高村 2006] 高村大也, 乾孝司, 奥村学, スピンモデルによる単語の感情極性抽出, 情報処理学会論文誌, Vol. 47, No. 2, pp. 627-637, 2006.
[Ohyama2007] K. Ohyama, T. Itoh, DIVA: An Automatic Music Arrangement Technique Based on Impressions of Images, Lecture Notes in Computer Science, Vol. 4569 (Smart Graphics 2007), pp. 178-181, 2007.
[Uno2014] A. Uno, T. Itoh, MALL: A Life Log Based Music Recommendation System and Portable Music Player, ACM Symposium on Applied Computing, Multimedia Visualization Track, pp. 939-944, 2014.
[Koyama2014] Y. Koyama, D. Sakamoto, T. Igarashi, Crowd-powered parameter analysis for visual design exploration, ACM Symposium on User Interface Software and Technology (UIST '14), 65-74, 2014.
[Koyama2016] Y. Koyama, D. Sakamoto, T. Igarashi, SelPh: Progressive Learning and Support of Manual Photo Color Enhancement, ACM Conference on Human Factors in Computing Systems (CHI '16), pp.2520-2532, 2016.
[斉藤 2016] 斉藤, 伊藤, CrowdRetouch: ユーザの画像補正傾向に基づく画像一括補正システム, 芸術科学会論文誌, Vol. 15, No. 4, pp. 147-156, 2016.