# 5 音楽と機械学習

吉井和佳（京都大学）

## 機械が音楽を理解する

機械学習は，音楽の解析・検索・生成などさまざまな研究課題を支える基盤技術である．たとえば，大規模な楽曲データベースから，ユーザの好みに合う楽曲を検索・推薦するためには，ジャンル，アーティスト名，人気度・試聴回数といったメタデータを利用して，各ユーザの好みを「学習」しなければならない．このとき，音楽音響信号から音楽的な特徴を自動的に抽出できるように「学習」しておくことは有益であろう．また，音楽を人間のように表情豊かに演奏するためには，楽譜（MIDI データ）から演奏時の強弱や時間的変動への写像を「学習」する必要がある．

本稿では，音楽の自動解析技術を中心に，音楽と機械学習のかかわりあいについて解説する．音楽の自動解析の主な課題として，和音推定，ビート解析（ビートトラッキング），自動採譜などが挙げられる．これらは本特集の記事「1. 音楽と信号処理」でも述べられているように，音楽信号処理の主要タスクであるが，多くの場合，機械学習技術が用いられる．人間は，知らない楽曲であったとしても，その楽曲のジャンルをある程度予測することができるが，このように，同じジャンルに属する楽曲が持つ普遍的な情報を抽出しておく（学習フェーズ）ことで，新たな楽曲に対応する（予測フェーズ）ような形態の機械学習を**教師あり学習**と呼ぶ．一方，人間は特別な音楽的訓練を受けていなくても（楽譜を読めなくても），楽曲を聴いただけで，音符というパーツが，ある規則に従って整然と配置されていることを直感的に把握することができる．この種の機械学習は，教師データ，すなわち音楽音響信号と楽譜データ（音符配置と音楽規則）の対応付けをあらかじめ覚え込ませておかなくても可能であると考えられ，**教師なし学習**と呼ばれている．

## 音楽解析における教師あり学習

本章では，ジャンル・ムード・印象認識，和音推定，ビート解析を題材に，音楽情報処理における教師あり学習の用いられ方を紹介する．

### ♫ ジャンル・ムード・印象認識

音楽音響信号をあらかじめ定められた複数のクラスのいずれかに分類することは，最も典型的な教師あり学習の問題であるため，長年さかんに取り組まれてきた．分類すべき「クラス」としては，**ジャンル**（genre），**ムード**（mood），**印象**（emotion）などいくつかのバリエーションが存在する．

印象認識では，各楽曲を 2 次元の Valence-Arousal 空間（VA 空間）内に位置付けることが一般的である．**図 -1** に示す通り，VA 空間は，横軸が Valence（Negative ↔ Positive），縦軸が Arousal（Silent ↔ Energetic）を示しており，これらの組合せでさまざまな印象を表現する．印象認識を行うには，2 次元空間内の座標を直接予測するアプローチと，図 -1 に示すように，空間をあらかじめ離散化しておき，いずれかのクラスに分類するアプローチが考えられる．前者のアプローチでは，楽曲全体を VA 空間内の一点に対応させる代わりに，楽曲を通した印象変化

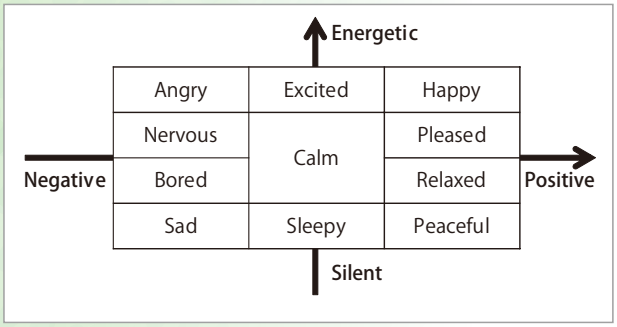


図 -1　Valence-Arousal 空間と印象ラベル

を VA 空間内の軌跡として表現する研究も存在する．後者のアプローチでは，クラスの粒度をどの程度に設定すべきかを十分に検討することが重要である．

通常，この種の分類タスクは，特徴量抽出と特徴量識別の 2 つのステップから構成されている．識別に利用する音響的特徴量としては，音声認識で用いられるメル周波数ケプストラム係数（MFCC）のほか，低レベル特徴量（spectral ux, spectral centroid, zero-crossing rate など）が用いられる．ただし，このような音響的特徴量は，短い時間フレームごとに得られるため，楽曲全体を表す特徴量とはなっていない．そのため，時間フレームの順序を無視することで，単なる特徴量の集合として捉える "Bag-of-Features" アプローチがよく採用される．得られた特徴量を識別するには，混合ガウスモデル（GMM），サポートベクトルマシン（SVM），ディープニューラルネットワーク（DNN）などが利用可能である．

### ♫ 和音推定

別の典型的な教師あり学習の問題が，本特集の記事「1. 音楽と信号処理」でも述べられていた**和音推定**である．ジャンル識別では，楽曲全体があるジャンルクラスに分類されていたのに対して，和音推定では，楽曲中の局所的な区間がコードクラスに分類される．認識対象となる和音の語彙はあらかじめ定めておく．たとえば，基本的な三和音のみ考慮する場合，ルート音 12 種類と和音タイプ major, minor の 2 種類の組合せに対し，無音などの非和音を表すシンボル "N" を加えた $2 \times 12 + 1 = 25$ クラスの分類問題となる．

ジャンル識別タスクと同様，和音推定タスクも特徴量抽出と特徴量識別の 2 つのステップから構成される．識別に利用する音響的特徴量として，12 次元のクロマベクトルが広く用いられている．クロマベクトルの要素は，一オクターブ内に存在する 12 個の音階 C, C#, …, B それぞれの強度を表す．コード認識では特徴量の時間的順序に意味があるため，識別器には隠れマルコフモデル（HMM）を用いることが一般的である．HMM の潜在変数系列が和音系列に対応しており，潜在変数が切り替わるところで和音が切り替わる．一方，各和音に対するクロマベクトルの分布は GMM や混合 von Mises-Fisher モデルで表現できる．近年は，音声認識システムと同様に，和音推定においても DNN の利用が進められている．また，スペクトル系列を入力し，和音系列を直接出力するような再帰型ニューラルネットワーク（RNN）を学習する試みもなされている [1]．

### ♫ ビート解析（ビートトラッキング）

**ビート解析**（ビートトラッキング）とは，音響信号に対してビート（拍）や小節線の位置などを検出・推定する問題である．詳細は本特集の記事「1. 音楽と信号処理」に譲るが，これまで信号処理的な手法やヒューリスティクスに頼っているケースが多かった．近年は信号処理技術のみに頼った手法から脱却し，RNN を教師あり学習する手法が提案されてきている [2]．

## *音楽解析における教師なし学習*

機械学習を用いた音楽情報処理タスクの花形は，**自動採譜**であろう．音響信号からその演奏内容を楽譜にする自動採譜は，本特集の記事「3. 音楽と音声情報処理」でも述べられているように，音声認識と並んで「音に関する夢の技術」といってもよい．

自動採譜は，近年の統計的機械学習の利用によって目覚ましい発展を遂げている．音楽音響信号は，楽譜を演奏することで生成されたものであるので，その生成過程を適切に記述することができれば，音

楽音響信号が与えられた際に，その背後にある楽譜を推定することが可能になる．このとき，演奏自体のゆらぎや，推論を行う上での不確実性を適切に取り扱うためには，確率的生成モデルを定式化することが望ましい．

音楽音響信号に対する自動採譜を行う上では，非負値行列分解（NMF）を利用することが一般的である[3]．詳細は本特集の記事「1. 音楽と信号処理」をご覧いただきたいが，混合音の振幅スペクトログラムのみから，（教師データなしに）基底スペクトル（ある楽器のある音高の平均的なスペクトル）の集合とそれらの音量の集合の両方を求めることができる点が特徴的である．最終的に，音量に対して閾値処理を行えば楽譜に準じる表現であるピアノロールが得られる．調整が困難な閾値処理を回避するため，各時刻で各音高が発音しているかを示す 2 値変数を NMF に組み入れる試みも行われている．しかし，得られるピアノロール自体の良さを評価する仕組みがないため，音楽として不自然な音高配置が頻発する問題があった．

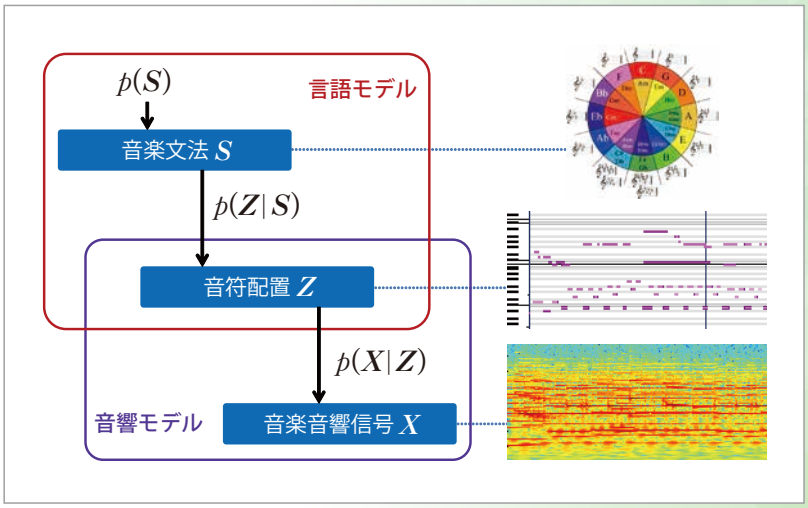

図 -2　自動採譜のための言語モデルと音響モデルを統合した階層ベイズモデル

最近，推定すべき楽譜自体の生成過程を確率モデルで表現することで，楽譜の生成モデルを事前分布，音響信号の生成モデルとを尤度関数とみなし，両者をベイズ的に統合する先駆的な試みが注目されつつある（**図 -2**）．これは，標準的な音声認識システムにおける**言語モデル**と**音響モデル**の統合と同型であり，採譜結果を音楽的に妥当なものに誘導する効果がある．ただし，自動採譜では，両者を一挙に教師なし学習する点で異なる．具体的には，まず，言語モデルと音響モデルを適当に初期化し，両者を考慮しながらいったん採譜を行う．採譜結果があれば，言語モデルと音響モデルをそれぞれ独立に更新できる．その結果，さらに良い自動採譜が行えるようになる．このような反復最適化は，音符配置を推定すると同時に，その背後にある音楽文法（言語モデル）を教師なしで導出しようとする野心的な試みである．

楽譜（音符配置）に対する言語モデルとしては，HMM を用いてコード（音符の組合せ）の遷移をモデル化しようとするものや[4]，確率的文脈自由文法（PCFG）を用いて，音符の背後にある階層構造を推定しようとするものがある[5), 6)]．特に，後者については，計算機上で音楽を取り扱うための理論である Generative Theory of Tonal Music（GTTM）を，生成モデルの見地から再解釈したものと考えることができる．

## 今後の展望

本稿では，音楽情報処理分野における機械学習，特に教師あり学習と教師なし学習の利用法について，ジャンル・ムード・印象認識，和音推定，ビート解析，自動採譜を題材に解説した．教師あり学習は，教師データを適切に用意することで，有用な識別器や分類器を容易に構築できるのがメリットである．一方，我々人間は音楽の文法や和音などの知識を教えられなくても音楽を楽しむことができる．そういう点では，音楽の音響モデルと言語モデルをすべて教師なし学習で構築する試みは，きわめて自然なアプローチに思える．これまでのところ，教師なし学習にはベイズモデリングが，教師あり学習にはディープラーニングが強みを発揮しているが，これらの技術を

融合することにより計算機による音楽理解のさらなる飛躍が期待される．

「音楽と機械学習」の分野のもう 1 つ重要な側面が，音楽生成への応用である．従来の自動作編曲研究では，音楽理論をベースにルールセットをシステム開発者が定義・実装することが多かったが，機械学習の利用がさかんになりつつある．こちらについては紙面の制約により省略するが，本特集の記事「4. 音楽とコンテンツ生成」で少し触れられている．

音楽は楽譜という離散記号の集合で記述できることから，自然言語処理技術の導入により，音楽解析技術の飛躍的な発展が期待できる．自然言語処理における形態素解析・構文解析・機械翻訳などの課題は，和音の機能解析・音符の階層構造解析・編曲などに対応すると考えられる．これまで，音楽解析の研究は音響信号を対象とするものであっても，それが「音楽」音響信号である必然性は薄かった．音楽の本質は離散記号上の構造にこそあるが，その構造・規則は自然言語のようには明文化しにくい．それを機械学習を用いて明らかにしていくことは，学術的にも意義が深い．

**参考文献**


1）Boulanger-Lewandowski, N., Bengio, Y. and Vincent, P. : Audio Chord Recognition with Recurrentneural Networks, *ISMIR*, pp.127-133 (2013).
2）Böck, S., Krebs, F. and Widmer, G. : A Multimodel Approach to Beat Tracking Considering Heterogeneous Music Styles, *ISMIR*, pp.603-608 (2014).
3）Smaragdis, P. and Brown, J. C. : Non-negative Matrix Factorization for Polyphonic Music Transcription, *WASPAA*, pp.177-180 (2003).
4）Yoshii, K. and Goto, M. : A Vocabularyfree Innity-gram Model for Nonparametric Bayesian Chord Progression Analysis, *ISMIR*, pp. 645-650 (2011).
5）Kameoka, H., Ochiai, K., Nakano, M., Tsuchiya, M. and Sagayama, S. : Contextfree2D Tree Structure Model of Musical Notes for Bayesian Modeling of Polyphonic Spectrograms, *ISMIR*, pp.307-312 (2012).
6）Nakamura, E., Hamanaka, M., Hirata, K. and Yoshii, K. : Tree-Structured Probabilistic Model of Monophonic written Music based on the Generative Theory of Tonal Music, *ICASSP* (2016).


（2016 年 3 月 17 日受付）

**吉井和佳（正会員）** yoshii@kuis.kyoto-u.ac.jp

2008 年，京都大学大学院情報学研究科博士後期課程修了．同年，産業技術総合研究所に入所．2014 年，京都大学大学院情報学研究科講師に着任．博士（情報学）．