6 音響ベースの音楽信号分類

George Tzanetakis[1]　訳：角尾 衣未留[2]

[1] ヴィクトリア大学コンピュータサイエンス科
[2] 東京大学大学院情報理工学系研究科

自動ジャンル分類は間違いなく音響音楽情報検索（Music Information Retireval : MIR）の中で最も広く行われるタスクであり，過去 10 年にわたり多くの結果が論文で発表されている．本稿ではこれら出版物全体から音響信号のさまざまな音楽ジャンル分類の研究のサーベイを行い，提案されている特徴量やアルゴリズムについて説明し，評価指標やデータセットについても議論する．さらに基本的なジャンル分類のタスクが拡張した他の音響信号分類のタスクについてもいくつか紹介する．

自動音楽ジャンル分類の問題設定

自動音楽ジャンル分類における問題がどのようなものであるかということは容易に示すことができる．音楽楽曲のレコーディングが与えられ，音響信号からコンピュータ解析によってその曲がどのジャンルやスタイルに属するかを認識するという問題である．処理の前に，特に自動音楽ジャンル分類の分野において用いられている，「スタイル」と「ジャンル」について簡単に説明しておきたいと思う．音楽学者である Fanco Fabbri は「ジャンル」を「あるコミュニティによって何らかの理由や目的，基準に基づいて定められた音楽の種類」と定義し，「スタイル」を「個人や土地や時代に特徴的な音楽イベント特徴が繰り返す演奏」と定義した．ジャンルの形成や進化，グループ化，変化のプロセスは音楽学や文化学によって調査されている．しかしここで扱おうとしている自動分類問題においては「ジャンル」と「スタイル」の違いはそこまで重要でない．なぜならば，少なくともシステムから見ると「ジャンル」や「スタイル」はあらかじめ決められたリスト（もしくは階層構造）のカテゴリーに属する楽曲によって特徴づけられているだけであるからである．そのため，自動音楽ジャンル分類における問題は以下のように言い換えることができる．ジャンル名のリスト（もしくは階層構造）と音響の音楽楽曲のコレクションに対して，音響信号が与えられたときそれがどのジャンルに属するかを高い認識率で推定するシステム（分類器）を作ること．このような音響信号の自動分類問題は音楽情報検索の分野で近年興っている研究の 1 つで，ディジタル音楽信号からの情報検索におけるすべての側面を扱う研究である．

本稿では音楽音響分類の文献で提案されているさまざまなアルゴリズムやその中のステージ，デザイン選択についての要点を説明する．紙面が限られているため，このトピックで出版されているすべての研究については言及できないが，既存の研究の代表例をあげ共通点・相違点を強調するようにしたい．自動音楽ジャンル分類は音楽分類の標準的な例であるが，最近ではさらにその発展形やその他の新しい問題が提案されている．その例にはアーティスト認識，自動ムード分類，自動タグ付け等があり，それらについても言及する．

一般的なシステム構造

自動音楽分類では一般的に以下のステージによって構成される．

- **正解情報の付与**　ジャンル分類を行う上でジャンルのラベルが付いた音楽トラックのセットがなくてはならない．さらに評価においてもラベル付きの音楽コレクションを用いる必要がある．ジャンルは変動的な概念でありラベルを付けることは見た目より簡単ではなく，これの同定はある程度の広がりを持っている．
- **音響特徴量抽出**　たとえば 1 秒に 44100 サンプルの音響信号など音響信号は多くの情報を持っており，音響信号の音楽コンテンツを直接理解するには十分すぎるほどである．このステージの目的は我々が音楽を聴くときに耳が行うのと似た方法で可能な限り音楽コンテンツを表す統計的な情報を抽出し集約することである．

- **楽曲の表現と分類**　このステージでは後に続く分類のためにどのようにして音楽楽曲を表現するかが重要なデザイン選択となる．しばしば既存の分類のための機械学習方法がこのステージで用いられる．しかし音響信号や音楽の特性に対応するために何らかの適応が必要である．
- **評価**　自動ジャンル分類が音楽情報検索の研究で人気のあるトピックである理由の1つに比較的単純に定量的な評価が行われることが挙げられる．それに対し自動音楽推奨問題はもっと評価が難しい．自動ジャンル分類では分類学や情報検索でのさまざまな評価指標が用いられている．また分類の性能を向上させるために少し音楽に特化したものが用いられることもある．また，もう1つ重要な問題は異なるシステムの結果を確実に比較できるようにするために共通のデータセットが手に入るかということがある．

文献で提案されている音響分類システムの相違点はこれらの各ステージでのアルゴリズムとデザイン選択に現れるといえる．次のいくつかの章で各ステージについてより詳しく述べると同時に関連する文献から具体的な例を挙げる．音響分類は音声認識の分野では長い歴史があるが，音楽に適用されたのはほんの最近である．いくつか初期の研究が発表されているが[1]，音響ベースの音楽ジャンル認識の良いスタート地点となるものは2002年にTzanetakisによって提案されたシステムであろう[2]．このシステムは一般的な構造とこの章で示したステージを実現し世界で初めて大量のデータコレクションに適用された(10ジャンルの1000曲)．

近年では自動ジャンル分類の研究の性能の進歩は次第に小さくなっている．そのため著者の中には分類性能に「ガラスの天井(glass ceiling)」があるかのような感覚を抱く者もあり，この問題は音楽ジャンルのあいまい性のため，これを目的にすることは実用的に限界があるとまで言われる．2006年にMcKayがなぜそれでもジャンル分類を行う価値があるか，どのように改善できるか納得のいく説明をしており[3]，さらに彼は文献のある項目により広く他の分類問題も含めてそれについて言及している．このトピックに対する興味が続く理由は，ユーザがジャンルを音楽情報の検索手段として頻繁に用いていることが研究によって示されているからであり，ジャンル分類には確立した評価方法があるため他のシステムと比較できる点にある．このトピックのより古いサーベイは文献4)で発表された．

音楽情報検索一般の研究や特に音響ベースの分類の研究の実験に無料で利用可能であるツールキットソフトウェアがいくつかあるので紹介する．Marsyas[☆1]やMIR Toolbox[☆2]，Music Analysis Toolbox[☆3]，機械学習用にWeka[☆4]等がある．

正解情報の付与

ジャンルは変動的な概念でそれらの精密な境界線を引くことは容易ではない．最も一般的なアプローチはAmazon[☆5]やAll Music Guide[☆6]等の大きな組織の情報源によるラベルを単純に用いるやり方である．このような情報源が持つ考えられ得る問題はジャンルが個々の曲に付けられるよりアーティストやアルバムごとに付けられがちであることである．そのため複数のジャンルに及ぶアーティストに付けられるラベルが誤りであることがあり，さらにあるアーティストに割り振られたジャンルと実際のジャンル分類が合致しているとも限らない．

音楽楽曲をあるジャンルへラベル付けすることはある程度主観的なことで，ジャンルの厳密な規則や定義があるわけではない．このことによって遂には音楽情報検索のために厳密に定義されたジャンルの階層構造を創ろうと提案する研究者まで現れることになったが，この提案は支持されることはなかった．

人間のジャンル認識のプロセスをより理解するためにGjerdingen[5]はある実験を行った．そこでは被験者に商用の音楽録音から抜粋された部分を聴き，大別した10の音楽ジャンルに分類してもらう．一般に与えられている正解情報と誰も完璧に合うことがないことがこの実験で示され（最高でも70%の合致），ジャンルの主観的な特性を示しより正確に自動分類システムの性能と人間の分類とを比べることができるようになった．もう1点この実験の興味深い結果は，被験者がたとえ4分の1秒の長さの抜粋に対してもそれなりの正解を示したことである．ほとんどの既存のジャンル分類システムにおいても計算コストを削減するために音楽楽曲のほんの一部（大抵10から30秒程度）が用いられ，分類性能には影響しないことが示されている．

上に挙げたような機関の情報源を用いずに正解情報を得るためのより踏み込んだ方法は複数の被験者にある決められたジャンルの中から曲にラベルを付けてもらうやり方である[6]．ある曲のジャンルラベルはその曲に対して被験者による多数決と同じ方法で与えられる．この結果は自動分類アルゴリズムはこのような大多数の意見に

☆1 http://marsyas.sness.net
☆2 http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/
☆3 http://pampalk.at/ma/
☆4 http://www.cs.waikato.ac.nz/ml/weka/
☆5 http://www.amazon.com
☆6 http://www.allmusic.com

よる正解情報を用いた時とこれに最も合致しなかった被験者によるラベルを用いた時とでほとんど同じような性能を示した．そして異なる意見を持つ人による異なるグループを配慮したより洗礼されたアプローチについて言及され，大多数意見の正解情報を用いる代わりに，被験者を共通の考え持つグループに分けそれぞれのラベルを別々に用いることが提案された．これは結局同じ曲でもユーザごとに異なるジャンルを予測するシステムを構築しようということになる．しかし残念ながらこのようなアプローチは大規模の主観実験が必要でありこの時点では提案だけがなされた．

音響特徴量抽出

音響特徴量抽出とは音響信号に潜んでいる音楽情報をまとめ簡潔に表現した値を計算することである．音響特徴量抽出は分類やセグメンテーション，類似検索など音響ベースの MIR のタスクのほとんどの基本となっている．この表現は人間がジャンルやスタイルのラベルを付けることを目的に音楽を聴いたとき人間が用いるさまざまな種類の音楽情報を統計的に捉えていなくてはならない．今のところ示されている音楽情報の 3 つの最も基本的なアスペクトは音色，リズム，和声である．音楽理論の分野でいまだ議題に上がっているようなこれらの厳格な定義よりも，音楽理論の知識が不必要であるような簡単な用語で説明したいと思う．音色は実際の演奏される音符に無関係にその楽器や音に関係する音楽の音の特性のことをいう．たとえばまったく同じ音楽楽曲をエレキギターやドラムでロックバンドが演奏したものとジャズビッグバンドで演奏したものとは音色はまったく異なっている．リズムはどの楽器が演奏されているかにかかわらず音楽に潜在する周期的に繰り返す構造のことをいう．たとえば有名なベートーベンの交響曲第 5 番の冒頭部はオーケストラシンフォニーで演奏されようと安いおもちゃのピアノで演奏されようと同じリズムである．和声は同時に鳴る別々の音高・音符のグループおよびそれを時間方向に展開したものをいう．たとえばビートルズのダンスリミックスは原曲と同じ和声構造をしているがリズムと音色は完全に異なっているであろう．

このような音楽情報の概念的に別々に異なる側面はお互いを影響し合っており，それぞれの音響特徴量は楽譜のようには正確とまでいかない統計的な近似しか与えないがそれでも分類器を学習する（分類システムを構築する）十分な情報は含んでいると考えられる．

音色情報は最も広く用いられている音響特徴量であり，今の時点ではそれ自体で用いた場合最も分類性能が良い．さらにこの特徴量は音楽以外の種類の音響にも適用することができ，むしろ音声認識の分野の方が長い歴史がある．このようにしてこの既存の音響特徴量が少し音楽へ適用させた形で用いられることもある．

音色特徴量抽出はさまざまな方法があるが，大抵のシステムは一般的に共通のプロセスで行われる．まず音響信号が短時間の断片に分けられ離散フーリエ変換等のフーリエ解析が行われ，その後数値のセット（特徴ベクトル）が計算されるデータ集約ステップを経る．この特徴ベクトルはこの短時間の断片のコンテンツ情報を要約し，捉えようとしたものである．このステージで音楽楽曲は高次元の空間で特徴ベクトル（点）の系列（軌跡）として表される．この系列はさらにまとまった表現を用いて表され，後に分類へ用いられる．

ほとんどの音響特徴量は 3 つのステージで抽出される．1）スペクトル計算，2）周波数領域での集約，3）時間領域での集約，である．スペクトルは周波数領域での音響信号のエネルギー分布であり，これの計算では短時間の区間（大概 10 から 40 ミリ秒程度）の波形サンプルを周波数領域表現へ変換する．このような変換の最も一般的なものは短時間フーリエ変換（Short Time Fourie Transform : STFT）である．各短時間区間ではおよそ定常であると仮定し，フレームの開始点と終点が不連続になる影響をなくすため窓関数がかけられる．この周波数領域への変換は信号の情報をすべて保存しており（逆変換が可能である），そのため変換の結果得られるスペクトルは高次元のものである．分析のためには，所望のコンテンツ情報は残しながらも著しく次元の小さい簡潔な表現にする必要がある．高次元のスペクトル（512 次元や 1024 次元の係数であることが多い）は小さい数の特徴量（10 次元から 30 次元程度）に集約されることが多い．よくある方法はスペクトル重心や帯域幅などスペクトルの形のさまざまな記述を用いることである [2]．その他広く用いられる周波数領域でのスペクトルの集約方法はメル周波数ケプストラム係数（Mel-Frequency Cepstral Coefficients : MFCCs）であり，会話認識・音声認識で生まれたものである．MFCC はスペクトルの情報（周波数ごとのエネルギー分布）を人間の聴覚機構の特性を考慮しながら集約するものである．このような特徴量は楽曲の音色に依存しており，人間が音色情報を知覚するのと同じように音色の「テクスチャ」が時間が進むにつれてどのように変化するかを表す．時間領域で集約する目的は短時間の分析区間よりもっと長い時間の信号を表現することである．この集約方法はしばしばおよそ 2，3 秒のいわゆる「テクスチャ」窓を用いて行われたり，1 曲全体に対して行われる．図 -1 に周波数と時間の集約を行った特徴量抽出を示す．

この特徴量集約の方法はいくつか提案されており，最

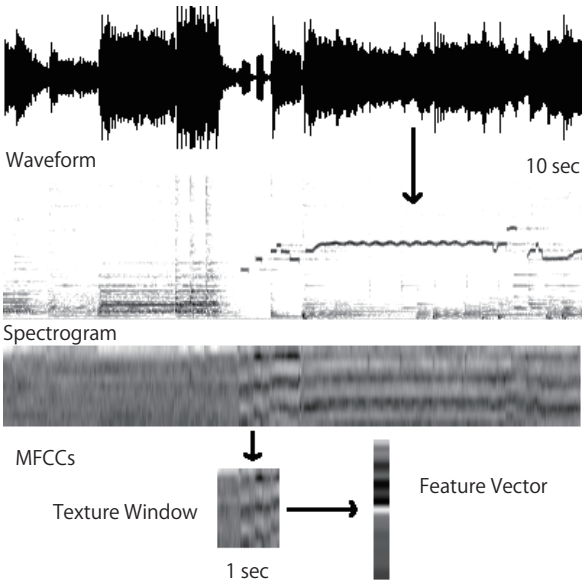

図 -1　特徴量抽出とテクスチャ窓

もよく用いられる手法は対角共分散や全共分散を用いたガウス分布関数をフィッティングしそのパラメータを特徴量ベクトルとする方法である（いわゆる平均や分散を求める方法）．その他のアプローチは自己回帰モデルを用いたものがある．

自動採譜は音楽信号を楽譜（音高とリズムの情報のみを持つシンボリックな表現方法）へ変換する処理である．これは難しい問題で既存の技術では単純な「おもちゃのような」例しか扱うことができない．その代わり広く用いられている音高を表す表現は音高と音高クラスの統計値（専門的に用いられる別名はそれぞれピッチヒストグラムとクロマベクトル）である．音高統計量は音楽の断片に存在する離散的な音高の出現率を計算したもので，音高クラス統計量はすべてのオクターブを本質的に同質として音高統計量を 12 の音高クラスへと折り畳んだものである．紙面の都合上どのようにして音高統計量を計算するかについては詳細には説明できないが，概して言うとある音高に対応するそれぞれの周波数のエネルギーを足し合わせるか，その代わりに複数の音高を同時に推定し，それをもとに統計量が計算される．

リズムの自動的な情報抽出もまた重要である．リズム情報は階層的な特性を持っており複数の関係する周期性を同時に含んでいる．典型的なものはビートヒストグラム（もしくは時にビートスペクトルとも呼ばれる）であらゆるテンポの可能性の中の主要点を表している [2]．また最新のものには自動的にあるジャンルを特徴付ける小節単位のリズムパターンを認識する手法がある．

現在の多くのポップやロックのレコーディングではそれぞれの楽器が別々に録音され最終的なミックスではレコーディングプロデューサ・エンジニアがリバーブやフィルタリングを加えステレオのパンニングキューを用いて左右のトラックに音響的な効果を加える．たとえば古いレコーディングでは生演奏のまま音源位置や左右の音のずれを保存し再現しているのに対し，最近では生演奏のセットではとても実現できないような効果を加えている．近年音響分類にステレオのパンニング特徴量も用いられている．その他の音響分類における興味深い可能性には音響ベースの特徴量を Web 上にあるレビュー文章等のその他の情報源との組合せがある．

楽曲の表現と分類

音響特徴量が抽出されればそれらは教師あり学習という技術によって分類器を「訓練」するのに用いる必要がある．訓練は訓練用の楽曲コレクションのすべてのラベルの付いた特徴量を用いて達成される．もし音響特徴量が 1 つの楽曲全体を表す高次元の特徴ベクトルに集約されているならばこれは分類の典型的な形となり，一般的に開発されているどの分類器でも利用することが可能である．音響ベースの音楽分類に対して用いられいる分類器の例は，ガウス混合モデル（Gaussian Mixture Models : GMMs）やサポートベクトルマシン（Support Vector Machines : SVMs），アダブースト（AdaBoost）等がある．これに代わる手法の 1 つにより短時間の区間に対して分類を行い多数決を用いて統計をとるやり方がある．より複雑な手法（よく bag-of-frames と呼ばれているが）では EM アルゴリズムという反復推定を用いたガウス混合モデルの学習のような分布推定手法を用いてそれぞれの楽曲をモデル化している．この場合各楽曲は高次元の 1 つの点（特徴ベクトル）ではなく 1 つの確立分布に相当する．確立分布間の距離は確立密度推定に用いられるパラメータ項に基づいたたとえば KL ダイバージェンスやそれの近似であるたとえばモンテカルロ法を用いて推定することができる．楽曲間の距離測定法を確立することによって k 近傍法（k-nearest neighbors : k-NN）などの単純な手法を用いることにより検索や分類を行うことが可能となる．

評価

分類の評価は比較的単純なものであり，大抵の場合どのような分類タスクとも同じである．標準的な手法は推定されたラベルをあらかじめ与えた正解ラベルと比較するやり方である．評価指標は検索での指標と同じように適合率（precision），再現率（recall），F 値（f-measure）のように共通のものがある．検索の指標が用いられる場合にはあるクエリ（検索語・楽曲）に対して関係性のある文章と同じジャンルのラベルが付いた楽曲とが対応すること

が前提とされている．交差検定は分類を評価するときに頻繁に用いられ，それはラベルの付いたデータは訓練用と試験用とに別々に分け，その訓練用と試験用に分割する分け方によって影響を受けないことを保証する方法である．1つ考慮に入れなくてはならない点はいわゆるアルバム効果（albumeffect）であり，訓練用と試験用のデータ両方に同じアルバムの楽曲が含まれている場合分類性能が不当に良くなってしまう効果のことである．一般的にはこれに対応するため訓練用もしくは試験用のデータセットのどちらかのみに確実に同じアルバム・アーティストの楽曲が含まれるように交差検定を行う．

同じデータセットを用いて同じ交差検定で分類精度を評価することによって異なるアルゴリズムやデザイン選択による相対する性能同士を比較することができる．正解情報の付与の章で述べたジャンルの主観性のため認識率を絶対的なものとして分類精度を評価することは正当でない．音響ベースの音楽ジャンル分類の初期の研究ではそれぞれの研究者が異なるデータセットと交差検定手法，評価指標を用いていたため，異なるそれぞれの手法のメリットを導き出すことが難しかった．データセットを共有することは著作権の規制によって難しくなっている．音楽情報処理の評価コンテスト，The Music Information Retrieval Evaluation Exchange（MIREX ☆7）では異なるMIRのアルゴリズムが音響ベースの分類タスクのようなさまざまなタスクがさまざまな指標で毎年評価されている．参加者は自ら提案するアルゴリズムのみを提出することによって著作権の問題に触れずに共通のデータを用い，あるデータセットへの過学習（over-fitting）も避けることができる．表-1は2008年開催のMIREXでの異なる音響分類タスクの最も性能の高かった結果を示している．Audio Tag Classification以外のすべての結果は分類精度（%）である．音響タグ分類はパーセンテージの代わりに平均のF値を用いている．

さらにRWCデータベースのような著作権フリーで利用可能のデータセットも公開されている．

ジャンル分類の発展形

自動音楽ジャンル認識は音楽分類の種類の中で最も研究されているが，それ以外にも研究されているトピックがいくつかある．このようなジャンル分類の発展と他の新しい問題は音楽ジャンル分類と多く共通点が見られ，どのように異なるかだけに焦点を当てて紹介したい．自動アーティスト認識はどのアーティスト・グループの楽曲かのラベルが付いていない音響レコーディングが与えられたときにある決められた選択肢の中からどれに属するかを推定する問題である．特徴量抽出や分類器の訓練という観点からすればジャンル分類とまったく同じ問題であるが，この問題はより多くのクラスに分ける必要があり学習のためのそれぞれのクラスの事例も少ない．音楽情景やムードの分類もまた同じようなアルゴリズム設計で行われるが，より正解情報の定義が難しい．

☆7 http://www.music-ir.org/mirex/2008/index.php/Main_Page/

Genre Classification	66.41
Genre Classification（Latin）	65.17
Audio Mood Classification	58.2
Artist Identification	47.65
Classical Composer Identification	53.25
Audio Tag Classification	0.28

表-1 音響ベースの音楽信号分類タスク（MIREX 2008）

まとめと今後の展開

自動音楽ジャンル認識は過去10年で音楽情報検索の主要な研究項目であった．今までさまざまなシステムが提案されているが，いずれも音響特徴量を抽出し機械学習に用いるという組合せを用いた共通構造を持っていた．また，決して自明なことではないが，音響分類システムの評価は今や成熟し共通の尺度とデータセットを用いることで異なるシステムの意義ある比較を可能にしている．関連する新しい問題としては自動タグ付け等が今後の研究の興味の矛先として興っている．

参考文献

1) Lambrou, T., Kudumakis, P., Speller, R., Sandler, M. and Linnery, A. : Classification of Audio Signals Using Statistical Features on Time and Wavelet Transform Domains, in Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (1998).
2) Tzanetakis, G. and Cook, P. : Musical Genre Classification of Audio Signals, IEEE Transaction on Speech and Audio Processing, Vol.10, No.5, pp.293-302 (2002).
3) McKay, C. and Fujinaga, I. : Musical Genre Classification : Is It Worth Pursuing and How Can It be Improved ?, in Proc. Int. Conf. on Music Information Retrieval (2006).
4) Scaringella, N., Zoia, G. and Mlynek, D. : Automatic Genre Classification of Music Content : A Survey, IEEE Signal Processing Magazine, Vol.23, No.2, pp.133-141 (2006).
5) Gjerdingen, R. and Perrot, D. : Scanning the Dial : The Rapid Recognition of Musical Genre, Journal of New Music Research, Vol.37, No.2, pp.93-100 (2008).
6) Lippens, S., Martens, J. P., Leman, M., Baets, B., Meyer, H. and Tzanetakis, G. : A Comparison of Human and Automatic Musical Genre Classification, in Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2004).

（平成21年7月16日受付）

George Tzanetakis gtzan@cs.uvic.ca

2002年プリンストン大学コンピュータ科学博士課程修了．音楽ジャンル自動分類に関する論文は頻繁に引用され，2004年IEEE Signal Processing Society 若手著者賞を受賞．現在ヴィクトリア大学コンピュータ科学科助教授．

角尾 衣未留（学生会員） tsunoo@hil.t.u-tokyo.ac.jp

2008年東京大学工学部計数工学科システム情報工学コース卒業．現在，同大学院情報理工学系研究科システム情報学専攻修士課程学生．