3 音楽と音声情報処理

齋藤大輔（東京大学）

「音の情報処理」の兄弟

画像を対象とした情報処理は画像処理と呼ぶ．何をいまさらと考える読者諸氏もいらっしゃると思うが，それでは取り扱う対象となるメディアを変え，音に対して同様の類推を行うとすればどうだろうか．「音」は訓読みなので，さしずめ「音の情報処理」と呼べる．しかし，実際には対象とする音の違いから，音楽情報処理と音声情報処理は異なる情報処理として定義され，それぞれ独自に大きな研究分野をなしている．音声情報処理は人間の音声言語を介したコミュニケーションの理解と再現を究極の目的としているといえる．そのための要素技術として音声認識や音声合成が重要な研究として位置づけられ，活発に研究が行われている．一方，音楽情報処理について，やはり同様の類推を行えば，これは音楽を介したコミュニケーションの理解と再現が目的と捉えることができる．音楽に含まれる多様な要素に対して，たとえばコード認識，自動採譜，自動作曲，音楽加工などのさまざまな技術が研究されている[1)]．

これらの情報処理は，いわば「音の情報処理」の兄弟と呼べ，共通点・類似点の多い研究課題が相互に存在する．しかし一方で，それぞれの対象の性質やとりまく諸相の違いから，研究課題が異なるものも存在する．本稿では，音声情報処理における主要な研究課題を切り口に，理解のための技術，再現のための技術のそれぞれに対応する音楽情報処理における研究課題を俯瞰する（**図-1**）．また相違点を元に，それぞれの研究分野における新たな可能性についても示す．

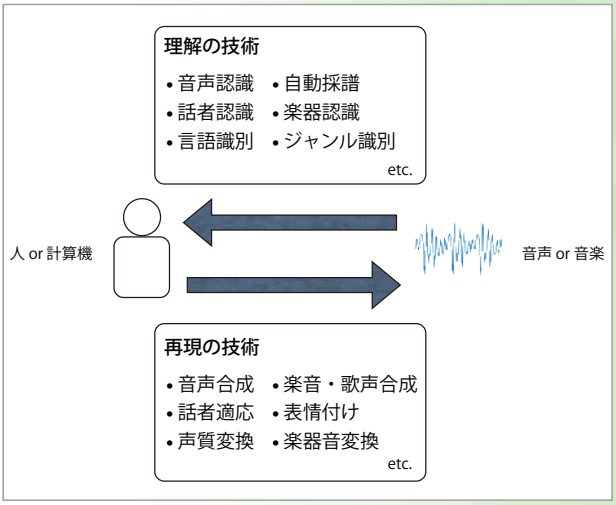

図-1　「音の情報処理」の俯瞰

理解のための技術

♫ 音声認識とその周辺

音声情報処理において，与えられた音声の発話から，その発話内容を知る技術が音声認識である．広義には発話内容の意図・意味にまで踏み込んだ言語理解を含むが，ここでは狭義の音声認識として与えられた音声の発話内容をテキスト情報に変換する技術（音声からの書き起こし：ディクテーション）について述べる．

標準的な音声認識では，与えられた音声特徴量（メルケプストラム係数：MFCC や対数メルフィルタバンク係数）と音素の対応をモデル化する**音響モデル**と，単語系列の並びの出現しやすさをモデル化する**言語モデル**をそれぞれ大規模なコーパスから学習し，これらの結果を**デコーダ**で統合することで最終的な音声認識結果であるテキスト列を出

力する．音響モデルとしては系列のモデル化に適した生成モデルである隠れマルコフモデル（HMM）が用いられるほか，近年では深層学習の発展に伴い，ディープニューラルネットワーク（DNN）とHMMのハイブリッドモデル（DNN-HMM）や，系列の依存を考慮したリカレントニューラルネットワーク（RNN）も用いられるようになった．言語モデルとしては直近の出現単語との依存を表すN-gramや，上述のRNNが用いられている．デコーダ部では，2つのモデルによって得られる認識仮説を，入出力シンボルと遷移確率を持つ有限状態機械である重み付き有限状態トランスデューサで表し，最終的な認識結果を得る．

なお音声言語を入力として，そこから情報を取り出す技術（狭義の理解技術）としてはほかに，誰の発話かを推定する話者認識，どの言語の発話かを識別する言語識別，発話者の感情状態を推定する感情・状態推定などが挙げられる．

♫ 自動採譜とその周辺

音楽情報処理における理解の技術で，音声情報処理における音声認識に相当するのが，与えられた音楽音響信号をその楽譜へと変換する自動採譜である．本特集の記事「5. 音楽と機械学習」にもある通り，音楽音響信号は一般に複数の音源の足し合わせであり，これを非負値行列分解などで適切に分離した上で，分離された音源の音高や音価などを推定することで最終的な楽譜を得るのが一般的な自動採譜の流れである．音声認識に比べると混合音である点が問題設定として大きく異なるため，前述のような音響モデルと言語モデルを統合するようなアプローチは，むしろ先進的な位置づけとなっている．また音声認識においては音響モデルの精緻さが比較的大きな比重を占めると言われているが，それと比べると高精度な自動採譜のためには，言語モデル，すなわち音楽的知識を考慮した音符配置に関するモデルの重要度が相対的に高いと考えられる．

また入力対象が歌唱音声の場合，基本的に音声情報処理と同様に歌唱音声認識，歌唱者認識を考えることができる．しかし楽譜に伴う発話長の違い，歌声と話声の違いなどによりその技術的な課題は異なってくる．

音声情報処理におけるその他の理解技術と，音楽情報処理との対応について考える．話者認識に対応したものはどの楽器から演奏されたかを推定する楽器認識，言語認識に対応したものはどのようなジャンルの音楽かを推定するジャンル推定が該当すると考えられる．演奏者認識は，音声情報処理における話者認識というよりも，演奏表情に基づいた推定となるため，感情・状態推定が該当すると考えられる．

再現のための技術

♫ 音声合成とその周辺

音声合成は，言語的な内容が与えられたときにそれを実際に発話された音声信号へと変換する技術である．狭義にはテキスト情報から音声への変換であるテキスト音声合成（Text-to-Speech：TTS）を考えるが，音声対話などでは概念からの発話テキストの生成（言語生成）を含む広義の概念音声合成（Concept-to-Speech：CTS）も検討されている．

TTSでは，一般に特定の話者の発話音声とそれに対応するテキスト情報が与えられたコーパスをもとに，テキストから音声への変換システムを構築する．主にTTSでは，音声波形を2つ組音素（ダイフォン）程度の時間単位で素片として保持しておき，入力テキストに合わせて適切な素片を接続する**素片接続型音声合成**と，メルケプストラムや基本周波数などの抽出された音声パラメータとテキストとの対応関係を統計モデルで学習し，合成時にはモデルから出力されたパラメータ系列から音声を生成する**統計的パラメトリック音声合成**がある．素片接続型音声合成は音声素片自体とテキストとの適合度を表す素片コストと，2つの素片のつながりやすさを表す接続コストをもとに，動的計画法によってコスト最小な素片系列を探索する問題として定式化される．一方，統計的パラメトリック音声合成では，尤度最大，または誤差が最小となるパラメータ系列をモデ

ルから出力する．このモデルとして，HMM のほか，近年では DNN を用いた手法も検討されている．

音声の再現に関連する技術として，既存の TTS モデルを少量の別の話者の音声を用いて，当該話者の TTS モデルへと適応する話者適応や，入力された音声について，その発話内容を維持しながら別の話者の声に変換する声質変換技術がある．

♫ 歌声合成・楽音合成

音楽を再現する技術において，個々の音源の自然性の向上は大きな課題である．楽音合成は，より自然な楽器音の合成を目的とし，音声情報処理の技術と対比すれば，高精度な分析合成技術に相当する．音声における分析合成がソースフィルタモデルに立脚したものであるのと比べ，楽音合成では，FM 音源，PCM 音源，そして物理モデルに立脚した音源の再現という流れで発展してきている．

再現の対象が歌声の場合，これは歌声合成と呼ばれる．音声合成における分類と同じくボーカロイド技術に代表される素片接続型と HMM 歌声合成に代表される統計的パラメトリック型を考えることができる．音声情報処理における音声合成や話者適応，声質変換に対応する研究も数多く行われている．歌唱と話声では，楽譜に伴う継続長や音高の制御，ビブラート等の話声にない特徴的な発声となるため，独自の研究視点を要する．

♫ 自動演奏表情付け

音楽情報処理における再現のための技術のうち，楽譜からの演奏は音声情報処理における音声合成と対応する．この場合 MIDI 音源等を用いることで演奏自体を自動で生成することはそれほど難しくない．しかし人間の演奏らしく聞こえる自動演奏のためには，演奏者ごとの演奏表情を付与する必要がある．このような演奏表情付け（Performance Rendering）は，国際的なコンペティションである Rencon が行われるなど，活発な研究が行われている．演奏表情付けは音声合成における話者適応と同様の位置づけと捉えることができるが，その目的が音声合成の話者適応の場合，柔軟性を目的としているのに対し，演奏表情付けの場合はより人間らしい演奏を目標とするなど，目的意識が異なっていることも興味深い．

新たな研究の可能性

本稿では，音声情報処理の主要な研究課題を紹介するとともに，音楽情報処理における研究課題との対応を紹介した．これらの研究分野はともに音を対象として，それを介した情報処理をさまざまな角度から行っていると解釈できる．今後，新しい研究を発展させる場合，このような対応を改めて検討することはさまざまな点で有用と考えられる．たとえば片方の分野で有効性が示されている技術を，スムーズにもう一方に導入する上で，このような問題定式化の共通点と相違点の理解が重要となる．現在これらの研究分野をともに研究対象としている研究者も多い．相互の技術理解を深めることで双方の研究分野がますます発展していくことが期待される．

参考文献

1）後藤真孝，平田圭二：音楽情報処理の最近の研究，日本音響学会誌 60 巻 11 号，pp.675-681(2004).

（2016 年 4 月 1 日受付）

齋藤大輔（正会員） dsk_saito@gavo.t.u-tokyo.ac.jp

東京大学大学院情報理工学系研究科助教．2011 年同大学院工学系研究科にて博士号（工学）取得，現在に至る．専門分野は音声合成を中心とする音声言語情報処理．