1 多重音解析と自動採譜

亀岡 弘和 [1]　　嵯峨山 茂樹 [2]

[1] NTT コミュニケーション科学基礎研究所
[2] 東京大学大学院情報理工学系研究科

自動採譜の効用

楽器経験者であれば誰しも自分の好きな曲を演奏してみたいと思うであろう．そして，好きな曲の楽譜がすぐに入手できたらどんなに便利だろうと一度は考えた経験があるものだ．もともとは，この要求に応えようとしたことが自動採譜の研究の始まりであった．しかしインターネット音楽配信サービスをはじめとする各種音楽関連のサービスの普及とともに，かつての演奏者支援という限定的な用途をはるかに超えて，自動採譜の研究の意義や効用が急速に拡大しつつある．

自動採譜とは文字通り，音楽の音響信号波形から人間の代わりにコンピュータを使って楽譜を書き起こさせるプロセスをいう．それでは楽譜とは何か．楽譜というと，一般のイメージでは図 -1 のような西洋音楽の記譜法による五線譜を思い浮かべるかもしれないが，たとえば，ギターの演奏では五線譜以外にタブ譜☆1 が，尺八の演奏では尺八譜が用いられたりするように，西洋式の五線譜に限らず演奏内容を描写する十分な情報を持った記号表現であればすべて楽譜なのである．そして，演奏するのは何も人間だけとは限らない．コンピュータも，MIDI☆2 メッセージ（図 -2 参照）のようにコンピュータが理解できる「言語」で書かれたデータをもとに，内蔵音源または外部音源を使って「演奏」することができる．したがって，MIDI メッセージもまた，演奏を再現するための演奏プロトコルという意味では楽譜の一種と言えよう．自動採譜とは，何らかの演奏プロトコルに従って生成された音響信号から，そのプロトコルを解読する逆プロセスなのである．

以上のように楽譜の概念を広義に捉えれば，楽譜とは人間だけでなくコンピュータが音楽を理解できるよう翻訳された「言語」であり，楽譜は演奏を再現する用途だけにとどまらず，使い道が実に多いことが分かってくる．たとえば，コンピュータの計算能力を活かしたさまざまな便利なアプリケーションが実現できるようになるのである．以下にその例を示そう．

☆1　各々の弦に対応した六線譜に数字で指板のポジションを示したもの．
☆2　Music Instrument Digital Interface の略．

■ 能動的音楽鑑賞

グラフィックイコライザくらいしか備わっていない従来の音楽プレイヤーでの音楽鑑賞では，既存の音楽を与えられたまま聴くことしかできず，リスナー側には音楽への働きかけの許されない受動的な聴き方が要求されていた．しかし，もし自動採譜が実現すれば，音楽鑑賞の仕方をより豊かにできる可能性がある．自動採譜により得られた楽譜を加工し，そのとおりに演奏を音響信号として再構成できれば，テンポを変え，楽器構成を変え，リズムを入れ替え，演奏スタイルを入れ替え，さらには編曲を変えるなど，リスナーの要求に合わせて音響信号

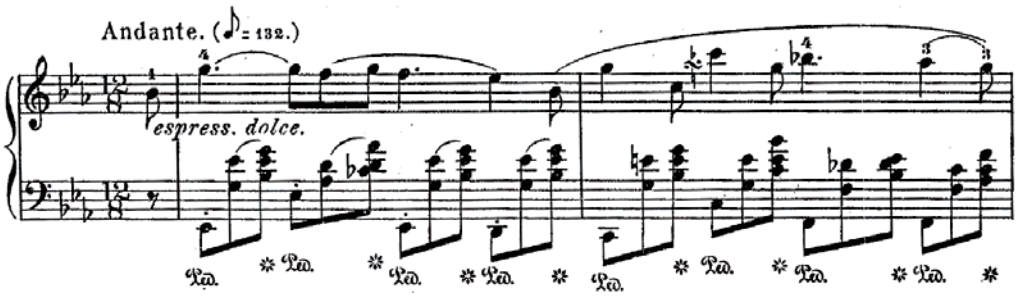

図 -1　楽譜例（F. Chopin 作曲 Nocturne, op.9, no.2）

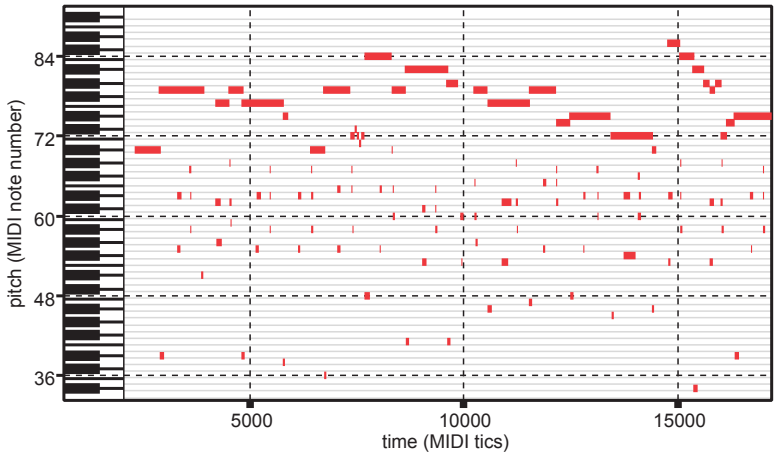

図 -2　図 -1 の MIDI メッセージ例．これも演奏を再現するための演奏プロトコルという意味で楽譜の一種である．

を柔軟にカスタマイズできる一歩踏みこんだ音楽鑑賞ができるようになるだろう．

■ CODEC

多くの楽曲においては，音階に対応する限られた種類の基本周波数の音だけで構成される，限られた種類の楽器の音だけで構成される，限られた種類の音長（八分音符，四分音符など）の音だけで構成される，限られた種類のメロディーやフレーズが時間を隔てて繰り返される，限られた種類のセクション（「A メロ」，「サビ」など）が繰り返される，といった特徴があり，さらに各音の波形は局所的に周期性を有することなどを踏まえると，音楽の波形データはきわめて高い冗長性を持ったデータと言える．したがって，MIDI メッセージのように波形レベルでなくノート単位の特徴パラメータを列挙したデータ形式は非常にデータ量が小さく済むという利点がある．このように，音楽信号の楽譜化は効率的な楽曲ストリーミングを可能にする CODEC☆3 の用途としても大いに役立つのである．

■ 音楽検索・著作権管理

インターネットを介した音楽の大量配信サービスの普及に伴い，リスナーの要求条件を満足するような楽曲との出会いを効率良く提供する音楽検索（推薦）システムの重要性が高まっている．音楽推薦システムについての詳しい解説は他の稿に委ねることにするが，多数のユーザの行動履歴を有効利用する現在主流の推薦方式とは別に，もしコンピュータが音楽を人間のように理解し，同等に感じることができるようになれば，これまで以上に潤滑に音楽との出会いを提供できる環境が整備できるはずである．また一方で，音楽コンテンツなどの著作物の使用状況をネットモニタリングする著作権管理技術の用途としても大いに活躍することになるだろう．このように音楽を人間と同等に理解し，感じることのできるコンピュータの実現のためには自動採譜技術の開発は避けては通れない．

自動採譜の難しさ

それでは，どのようにすれば自動採譜技術は実現できるのか．残念ながらその答えは今なお完全には分かっていない．

音楽の自動採譜は，さまざまな問題が複合的に絡んだ非常に大規模で複雑な問題であるがゆえに，問題解決の見通しを良くするためには比較的扱いやすい要素問題にうまく分解し，構造化し，整理することが重要である．そもそも人間はどのように問題をブレイクダウンして，音楽を理解しているのだろうか．人間がどのようなプロセスを経て音楽を認知しているかを知ることは問題解決への少なからずのヒントになるはずだ．

実は人間の脳内では，いくつもの機能単位に分割されたモジュール方式が採られているとの知見がある．図 -3 は，Peretz らによって脳損傷患者の機能障害に関する事例調査に基づいて明らかになった人間の聴覚認知における機能的アーキテクチャの一部を示している．この図によれば，音響信号はまず，acoustic analysis モジュールにより，音響信号が個々の音脈☆4 に分凝☆5 される．たとえば，複数の楽器音が同時に発音している場合には，形成されるそれぞれの音脈はほとんどの場合個々の楽器

☆3　Coder/Decoder の略である．

☆4　時間的にひとつながりに聴こえる音のかたまりを音脈という．

☆5　音響信号を細かい要素に分解し，個々の要素を組み立てて音脈を形成する聴覚のプロセスを分凝という．

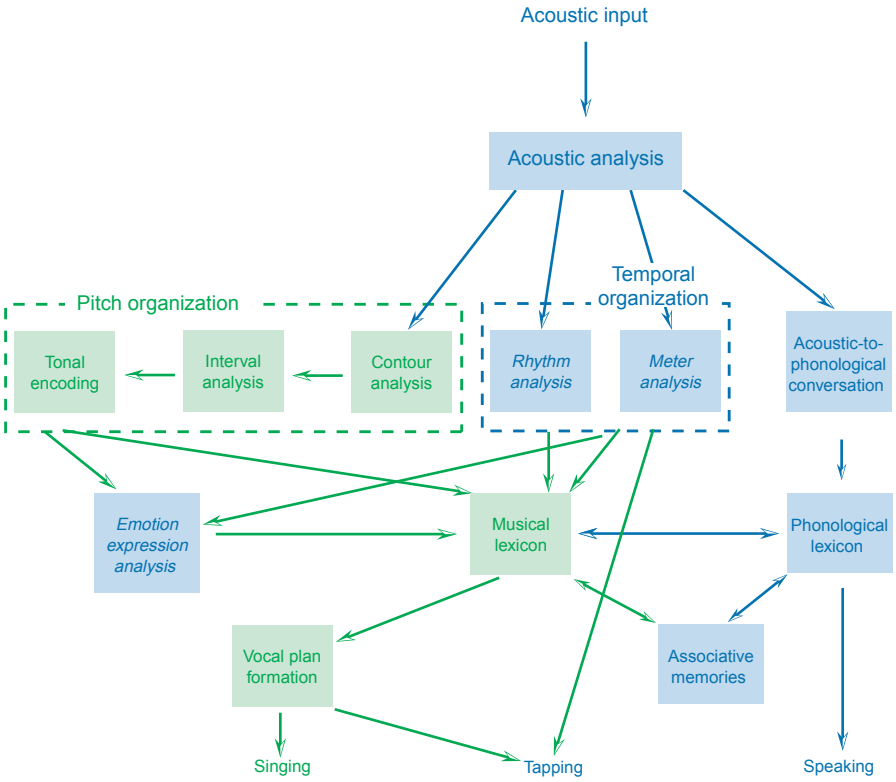

図 -3　Peretz らによって示された聴覚認知における脳内のモジュール方式（文献 1）より抜粋）．

音に対応する．このように混合信号から個々の音源に関する情報を抽出する処理を多重音解析という．そしてこのステップを経て，メロディー構造の認識とリズム・拍節構造の認識（pitch organization と temporal organization）がそれぞれ並行してなされ，その後さらに音楽フレーズ（lexicon）や音楽表現（emotion）といったより抽象化された情報に統合されていく．

本稿では，後続するすべてのモジュールに通ずる多重音解析の問題を中心に解説する．図 -3 の各モジュールに相当する要素問題すべてを本稿では議論しきれないため，リズム認識や拍節認識（ビートトラッキング）など，その他の要素問題に関する解説はたとえば文献 2），3）などを参照されたい．

多重音解析と音の群化

図 -4 は，多重音解析の難しさを端的に表している．これは，図 -1，2 の曲をピアノで演奏したときの音響信号を時間周波数解析したものであり，音響信号が各時刻でどのような周波数成分で構成されているかを表すスペクトログラムと呼ぶ信号表現である．この図を見ると，個々の打鍵音の周波数成分が複雑に重なり合っており，各時間周波数成分がもともとどの音に由来したものであるかを容易に知ることができない．

しかし興味深いことに，我々人間は，多数の音が混じり合った音響信号から，個々の音を難なく聴き分けることができる．足し算が不可逆であるのと同じように，いったん重畳されてしまった波形から個々の波形を復元することは一般には困難である．にもかかわらず，混じり合っている個々の音の「輪郭」や「境界」を正確に把握できるのは，人間の聴覚の「アルゴリズム」がいかに優秀であるかを示している．

この逆問題を，人間がどのようなアプローチにより解いているのかについては未解明な点が多い．両耳に入ってくる 2 つの波形の微妙な違いに基づいて知覚される波源位置の情報は，音を聴き分けるための手がかりの 1 つに違いないが，我々はモノラル録音された音響信号からですら個々の音を聴き分けられる能力を持っている．このことは，人間には空間的手がかり以外の手がかりに基づくなんらかの音の群化☆6 メカニズムが備わっていることを示唆し，この困難な逆問題を人間がどうにかして解いているという事実は，音を聴き分ける原理を追究することへの動機となっている．

ここでは，多重音解析の問題に焦点を当て，解くべき問題を明らかにしながらその計算論的アプローチに関する近年の取り組みついて紹介する．

☆6　人間は，物体を見たときに，どこまでをひとまとまりなのかを捉え，物体の輪郭を把握することができるのと同じように，音を聞いたときにも，どこまでをひとまとまりなのかを捉え，音の「輪郭」を把握することができる．このように，音の「輪郭」を把握することを「音の群化」といい，このように形成されたひとまとまりの音の「塊」のことを「音脈」という．

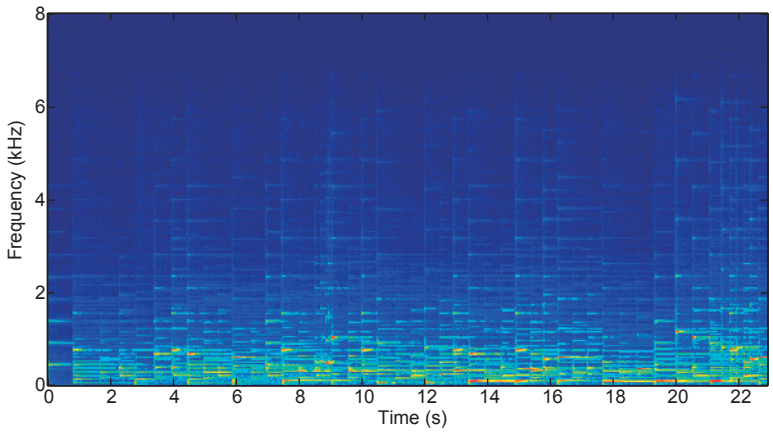

図 -4　図 -1，2 の曲に対応する音響信号のスペクトログラム

■ 基本周波数推定（周波数方向の群化）

音の群化の問題と多重音の基本周波数☆7 推定の問題との間には，きわめて密接な関係がある．このことを明確にするため，分かりやすい例題として単一音のパワースペクトルから基本周波数を推定する問題について考えよう．もし信号が純音の場合，パワースペクトルのピーク周波数が基本周波数に対応する（図 -5（a））が，一般の周期信号には複数のピークがある（図 -5（b））．そして複数あるピークのうち最大のピークの周波数が必ずしも基本周波数に対応するとは限らない（図 -5（c））．また，基本周波数成分はいつも大きいとは限らないため，複数あるピーク周波数のうち最も低い周波数を基本周波数と見なすのは頑健なやり方ではない（図 -5（d））．以上より，基本周波数を推定するためには，スペクトルピークのような限られた情報だけで済ませようとするのではなく，対象とする音の信号波形やスペクトル構造の全体を手がかりにしたロバストな方法が必要になる（単一音の基本周波数推定すら容易でないことは長い研究の歴史が物語っている [4]）．しかしながら，複数の信号が混合されて観測される音響信号には，どの成分がどの音に帰属するのかという情報が欠落しているため，基本周波数を推定するための重要な手がかりが得られないのである．したがって，音の群化の問題が解かれない限り，個々の基本周波数を推定することは容易ではないわけである．一方で，もし，個々の音の基本周波数が既知であれば（きわめて特異な状況であるが），各音に由来する成分の検討がつくため，音の群化の問題は大幅に解きやすくなる．すなわち，音の群化の問題を解く手がかりになる基本周波数の情報が，音の群化が解かれない限り安定的に求められない，といういわゆる「鶏と卵」の状況に陥るのである．

☆7 基本周波数とは，周期信号を構成する周波数成分の中で最も低い周波数であり，F_0 と呼ぶことがある．

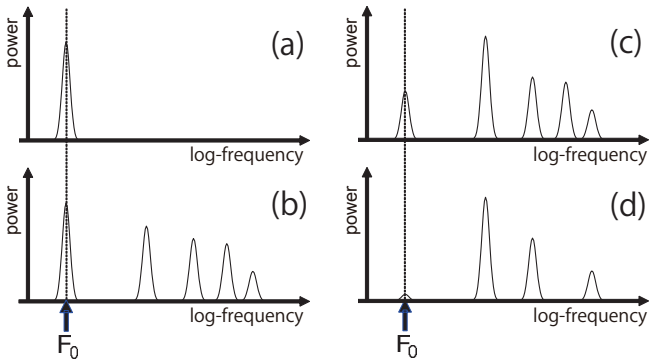

図 -5　基本周波数推定の問題

この問題に対しては，音に関する先験的知識（調波性やスペクトル概形の仮定）を利用する方法が主流な常套手段となる．たとえば，観測音響信号をあらゆる基本周波数の音の重みつき混合としてモデル化してその重みを推定するアプローチ，スペクトルクラスタモデルを用いて音の群化と基本周波数推定を反復的に行うスタイルのアプローチや対数周波数領域で調波構造のシフト不変性を仮定して対数周波数スペクトルを調波構造パターンで逆畳み込みするアプローチなどが試みられている．このほかにも，多重音から基本周波数を推定する手法は膨大にあるので，より詳しい動向については他の著書 [5] を参照されたい．

■ 計算論的聴覚情景分析（時間方向の群化）

前節では，暗黙のうちに非常に短い時間区間における波形から個々の音に群化する問題について考えていた．我々人間でも，数十ミリ秒程度の混合信号から個々の音を聴き分けるのは必ずしも容易ではなく，容易に聴き分けるためにはある程度の信号の長さが必要になる．前節で考えていた問題は，周波数方向の群化と呼ぶものに相

当し，人間はそれだけでなく音の時間的な連なりを形成する時間方向の群化も同時に行っているとされる．

近年，聴覚情景分析[6]と呼ぶ心理学的アプローチの枠組みによって徐々に明らかになってきた人間の音の群化メカニズムに関する知見を積極的に利用して，音の群化問題の解決を図ろうとする試みが進められており，その枠組みを総称して計算論的聴覚情景分析（Computational Auditory Scene Analysis : CASA）と呼ぶ．具体的には，知識を利用しない聴覚の低次の音の分離能力に関して，音響信号はスペクトログラムに似た要素に「分解」されること，同じ音源に由来する要素は「群化」されて音脈を形成すること，群化のされやすさ（分凝要件）は，(1) 調波性，(2) 調波成分の立ち上りの共通性，(3) 調波成分の周波数および振幅変化の共通性，(4) 成分の連続性，(5) 時間周波数の近接性，(6) 音源位置の共通性などに関係する，ことなどが心理実験を通して示されている．瞬時瞬時において調波的関係にある周波数成分を 1 つの音としてグルーピングすることを周波数方向の群化といい，それらを分凝要件（2）～（5）に基づいて継時的にグルーピングすることを時間方向の群化という．これによって，たとえば，2 つの音声の基本周波数軌跡がある時点で交差していたとしても，本来は分離不能なはずの交差の瞬間における個々の音声信号の各周波数成分がどのように重なっているかを前後の時刻から推論できるようになるわけである．CASA の目的は，このような人間の低次機能による音の群化メカニズムを模倣することであり，上記の分解と群化のプロセスを，分凝要件に関係する物理量を用いてアルゴリズムとして実現し，音脈の認識に有用な特徴量(基本周波数など)を抽出したり，目的音に相当する音脈の再構成を行うことである．

その具体的なアプローチとしては，周波数方向の群化に相当する処理により各離散時刻において個々の構成音の瞬時特徴成分(たとえばスペクトルや基本周波数)を抽出したのちに，マルチエージェントシステムやベイジアンネットワークや隠れマルコフモデルや Kalman フィルタなどの手段を通して，時間的にどの成分が同じ一連の音に対応しているかを瞬時特徴成分の時間的滑らかさなどを評価尺度にして推定する方法が主流である．また一方で，分凝要件(1) ～ (5)から逸脱しない範囲の自由度を持った時変スペクトルを直接的にモデル化し，これを混合したもので観測時間周波数スペクトルにフィッティングする，周波数方向および時間方向の群化を同時最適化問題として定式化されたアプローチも考案されている．

■スパース表現（記憶に基づく群化）

ところで我々は，ユニゾン（同一音高またはオクターブ違い）で弾かれたピアノとヴァイオリンの音を聴き分けることができる場合がある．一定の時間連続して一方の音の調波成分が完全に他方の音の調波成分と重なってしまうこの状況では，前後の時刻から調波成分の重なり具合を推論することが難しいため，これまで述べてきた群化メカニズムとは別の何らかのメカニズムが存在している可能性が示唆される．きわめてわずかな基本周波数の違いによって 2 つの信号の間に干渉が生じており，それを手がかりにしている可能性もあるが，それ以外に，ピアノやヴァイオリンがどのような音色であるかを漠然と記憶していて，それに基づいて個々の音脈を推論するような働きが関与しているとも考えられる．

たとえば，ピアノの音とヴァイオリンの音を過去にもっと容易に聴き分けやすい状況で聴き分けた経験があったとして，その経験から，それぞれの音響的特徴に関する「辞書」が作られているとすると，この「辞書」はユニゾンのような困難な状況においても高い精度で音を群化するための有用な手がかりになる．そして，このように音の群化が高い精度でなされるたびに，信頼性の高い学習データを得たことになり，「辞書」はより良く再学習される．

近年，スパース表現と呼ぶ枠組みを応用して，以上のような観点で音の群化の問題を捉えたアプローチが脚光を浴びている．このアプローチは，調波性など音源に関する仮定をほとんど置かない代わりに，音響信号の全体が限られた種類の要素だけで構成されていると見なし，その要素を情報論的な規準に基づいて推定しようというものである．具体的には，各時刻で観測される混合信号ないし混合スペクトルを，いくつかの構成パーツの重みつき和によってモデル化し，できるだけその重みをスパース☆8に，かつモデル化誤差を小さくするように各パーツの形状と重みを反復学習する．すると，観測信号中に繰り返し現れる信号あるいはスペクトルのパターンが次第に典型的な構成パーツとして見なされ，抽出される．通常，パーツと重み☆9は交互に更新されるため，ちょうど上述の例えと同様な反復学習が行われることになる．非負値行列分解(Non-negative Matrix Factorization ; NMF) は，パーツと重みをいずれも非負制約のもとで学習する方法であり，効率的な学習アルゴリズムが存在する点，非負制約以外の制約がなくとも副次的に重みがスパースになるよう学習される点が特徴的である．

☆8　重み係数のうち一部だけが大きい値をとり，それ以外はほぼ 0 であることを「重みがスパースである」という．

☆9　重みの更新は，前段で更新された信号あるいはスペクトルのパーツを使って観測信号にフィッティングさせる操作に相当するため，音の群化処理にほかならない．

自動採譜の今後

本稿では，図 -3 に示した人間の音楽認知プロセスのモデルに基づいて，自動採譜の問題をいくつかの小さい要素問題に分解し，その根幹となる多重音解析の問題を中心に議論した．図 -3 は，音響信号入力から高次の知識を抽出しようとする典型的なボトムアップ型の処理系を表しているが，このような一方向性の処理が自動採譜の問題への最適なアプローチになるとは限らない．これと対立するものとして，音楽理論に基づいて立てられる仮説的なモデルを音響信号に当てはめて，解釈しようとするトップダウン型のアプローチも考えられる．自動採譜の実現に向けて，いずれは，低次機能による個々の音への群化処理(多重音解析)と，それを音楽的制約に基づいて構造化して記号化する処理とを統合的に行える大規模な情報統合モデルの構築が必要になるであろう．

現在標準となっている連続音声認識システム（図 -6）は，大規模な情報統合モデルを確率モデルに基づいてエレガントに具現化した良い見本である．これを転用した仮想的な自動採譜システムを考えてみよう．図 -6 に示す連続音声認識システムにおいて，仮に音声信号入力を音楽信号入力に，音素モデルを音符モデルに，言語モデルを音楽モデルに置き換えれば，音声認識出力は自動採譜出力に置き換わる．ここで特徴抽出部では，多重音解析による各音の基本周波数とオンセット時刻の情報の抽出を行う．音符モデルは，ピッチ変動(ビブラートなど)，音長変動，音色変動を表す確率モデルである．音楽モデルには，リズム語彙，和声構造，旋律，楽曲構造，曲スタイルなど，音楽に関する常識すべてが該当し，それらをいかに確率モデルとして表現するかが課題となるだろう．そして，探索過程では，入力から抽出された特徴量を，音符モデルと音楽モデルの両方から最ももっともらしく説明できる原楽譜を効率良く探索することになる．

上記は連続音声認識のアプローチを踏襲した場合の自動採譜アプローチの一例であるが，自動採譜の問題を解くためには，要素問題を 1 つ 1 つ着実にクリアする精巧な視点と，上記の例のような巨視的な視点を併せ持ったバランスの良い解法を見つけていくことが今後の重要な課題となるであろう．

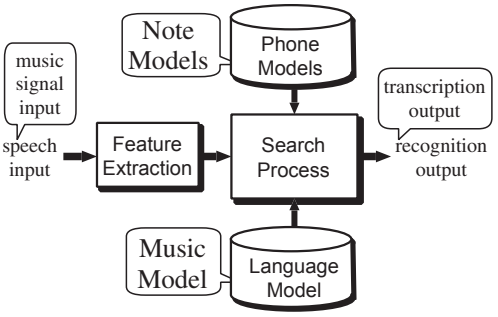

図 -6　連続音声認識システムの自動採譜への転用

参考文献

1）Peretz, I. and Coltheart, M. : Modularity of Music Processing, Nature Neuroscience, Vol.6, No.7, pp.688-691 (2003).
2）長嶋 , 橋本 , 平賀 , 平田 (編）: コンピュータと音楽の世界―基礎からフロンティアまで― , bit 別冊 , 共立出版 (1999).
3）Klapuri, A. and Davy, M. (Editors) : Signal Processing Methods for Music Transcription, Springer-Verlag, New York (2006).
4）Hess, W. : Pitch Determination of Speech Signals, Springer-Verlag, Berlin (1983).
5）de Cheveigné, A. : Multiple F0 Estimation, in Computational Auditory Scene Analysis : Principles, Algorithms and Applications, D. -L. Wang, G. J. Brown Eds., IEEE Press / Wisely (2006).
6）Bregman, A. S. : Auditory Scene Analysis, MIT Press, Cambridge (1990).

（平成 21 年 7 月 3 日受付）

亀岡 弘和（正会員）　kameoka@eye.brl.ntt.co.jp

平成 14 年東大・工・計数卒業．平成 16 年同大学院情報理工システム情報修士課程修了．平成 19 年同博士課程修了．情報理工学博士．同年日本電信電話（株）入社．以来，音声・音楽を対象とした音響信号処理の研究に従事．現在，NTT コミュニケーション科学基礎研究所勤務．第 20 回電気通信普及財団テレコムシステム技術学生賞，Best Student Paper Award Finalist at ICASSP2005，本会平成 17 年度山下記念研究賞，日本音響学会第 2 回独創研究奨励賞板倉記念，平成 18 年度東京大学大学院情報理工学系研究科研究科長賞，第 1 回 IEEE Signal Processing Society Japan Chapter Student Paper Award，日本音響学会第 25 回粟屋潔学術奨励賞，IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award を受賞．日本音響学会，電子情報通信学会，IEEE 各会員．

嵯峨山 茂樹（正会員）　sagayama@hil.t.u-tokyo.ac.jp

1974 年東京大学大学院工学系研究科修士課程修了．NTT，ATR 自動翻訳電話研究所，北陸先端科学技術大学院大学などを経て，現在，東京大学大学院情報理工学系研究科教授．博士（工学）．