音楽情報科学 7－9
（1994．8．7）

音楽音響信号から単音記号列を生成するシステム OPTIMA の全体像

柏野 邦夫　中臺 一博　田中 英彦
東京大学 工学部

kashino@MTL.t.u-tokyo.ac.jp

あらまし　知覚的音源分離における本質的な課題の一つは、対象に関する知識や記憶に基づく処理を柔軟に組み合わせて最終的な結果を求めることである。本稿では、仮説ネットワークによる階層的な情報統合と最尤推定のメカニズムを備えた、音楽音響信号を対象とする知覚的音源分離の処理モデル OPTIMA を提案する。この処理モデルは、複数種類の楽器音を含むモノラルの音楽音響信号をもとに楽器種類ごとの演奏情報を抽出して、単音記号列などの形で出力するシステムとして応用されている。本稿では処理モデルの全体像を示すとともに、特に情報統合の原理と仮説ネットワークの挙動について詳細に議論する。

OPTIMA : Organized Processing toward Intelligent Music Scene Analysis
– General Description of the Process Model –

Kunio Kashino　Kazuhiro Nakadai　Hidehiko Tanaka

H.Tanaka Lab., Bldg.#13, Department of Electrical Engineering,
Faculty of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-Ku, Tokyo 113, JAPAN.

Abstract　We describe OPTIMA, a process model for the perceptual sound source separation on computers. Our model consists of four parts: bottom-up processing modules, top-down processing modules, knowledge sources, and a hypothesis network for hierarchical and quantitative integration of multiple bits of information. First we present general description of the model. Since one of the most essential problems in the perceptual sound source separation is integration of multiple bits of information, we then focus our discussion on the hypothesis network: we show that our method has permitted efficient, autonomous and stable construction of an optimal internal model of the outer world.

1 まえがき

本稿では、知覚的音源分離システム [1] の処理モデルを新たに提案する。処理モデルの具体的な応用例は、複数種類の楽器音を含むモノラルの音楽音響信号を入力とし、楽器種類ごとの演奏情報を抽出して、単音記号列 (MIDI データ, 画面表示) および分離・再合成した各楽器ごとの音響信号として出力するシステムである。

これまでわれわれは、知覚的音源分離の考え方に基づき、モノラルの楽器演奏を対象とする音源分離システムについて検討を進めてきた [2, 3]。知覚的音源分離とは、人間がひとつのものとして知覚または認識するような音響エネルギーのまとまり (これを知覚的な音と呼ぶ) を一つのものとして記号化および構造化することを指す [4]。

知覚的音源分離は、混合された音の波形をもとに、外界の事象、即ち混合される前の音に対応する階層化された記号表現を求める問題であり、一般的な条件の下では不良設定問題となる (例えば、システムに入力される音の具体的なモデルが既知でなければ、重複周波数成分を分解できない)。これは、入力データに基づくボトムアップ処理だけでは処理精度に一定の限界があることを意味する。従って、対象に関する知識や記憶に基づく処理を柔軟に組み合わせて最終的な結果を求めることは、知覚的音源分離にとって本質的な課題である。

そこで本稿では、情報統合のメカニズムを備えた知覚的音源分離の処理モデル OPTIMA (Organized Processing toward Intelligent Music Scene Analysis) を新たに提案する。OPTIMA は、その名が表すように、各時点で得られた種々の情報に基づいて外界に対する最適 (optimal) な内部像を組み上げていく枠組である [1]。以下の各章において、処理モデルの具体的構成を示してその全体像を明らかにするとともに、各構成要素の機能について議論する。なお本稿は、特に、処理モデルの要となる情報統合の原理と仮説ネットワークの挙動を詳細に論じることを主眼とする。

2 OPTIMA の全体像

処理モデル (OPTIMA) の全体像を図 1 に示す。このシステムの入力はモノラルの音楽音響信号である。出力は楽器ごとに分類された単音の列である。

OPTIMA は、抽象度の低い順に (1) エネルギー表現、(2) 周波数成分、(3) 単音、および (4) 音楽シーンの四つの抽象度の階層を持っている。(1) エネルギー表現は、入力に含まれる音響エネルギーを時間–周波数平面上に表現したものである。(2) 周波数成分は、単音を構成する音響エネルギーの集まりであって、エネルギー表現において周波数方向のローカルピークを時間方向に接続することによって得られるものである。(3) 単音は、個々の音符に対応する記号表現である。

[1]ここでは、情報を得た後の事後確率が最大の仮説のセットを、最も尤もらしいという意味で「最適」と呼んでいる。

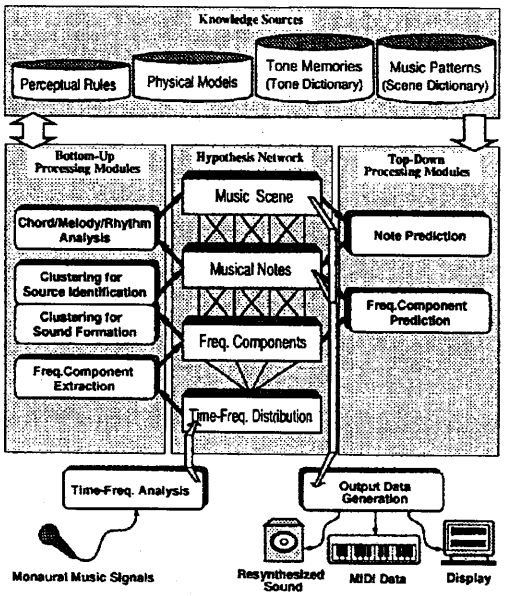

図 1: OPTIMA の全体像

(4) 音楽シーンは、ここでは複数の単音を特徴づける記号表現のことを指している。例えば、継時的な性質を持つ音楽シーンとしてはリズム、同時的な性質を持つ音楽シーンとしては和音が挙げられる。これらのうち、周波数成分、単音、シーンの三つの階層においては、システムは一般にそれぞれ複数の仮説を保持する。

OPTIMA は、四つのブロックから構成されている。四つのブロックとは、(A) 抽象度の低い階層から抽象度の高い階層への情報表現の変換を行うボトムアップ処理モジュール、(B) 抽象度の高い階層から抽象度の高い階層への情報表現の変換を行うトップダウン処理モジュール、(C) 処理モジュールに情報を渡したり処理モジュールから情報を受け取ったりする知識源、および (D) 情報統合のための仮説ネットワークである。以下で、それぞれの構成要素について議論する。

3 情報統合のモデル

3.1 情報統合メカニズムの要件

情報統合が知覚的音源分離システムにとって本質的な課題であることは既に述べたが、そのメカニズムはどのような特性を持つべきだろうか。

音声認識システムの研究を眺めてみると、これまでに階層モデル、黒板モデル、散層モデルなどの処理統合モデルが研究されてきた。そこでこれらのモデルを知覚的音源分離システムに用いることを検討してみると、従来の階層モデルは、処理の方向や順序が一定であることなど、システムとしての柔軟性に欠ける面があって使いにくい。黒板モデルは、柔軟な処理が可能である反面、処理の制御が困難である。また散層モデルは、強い制約となる確信度の高いデータ (音声認識システムにおける、対象単語のテンプレートなど) がある場合には有効であるが、知覚的音源分離システ

ムのようにひとつの抽象度階層においては対象に対する制約が緩く、全体としてある解釈の確信度が高まるといった性質の処理を行う用途には必ずしも適していない。

以上のことから、知覚的音源分離システムにおける情報統合のメカニズムとしては、

1. どのようなモジュールが仮説生成を行うかや、その順序に柔軟性がある
2. 処理の制御が容易である
3. 多くの緩い制約の積み重ねを効率的に扱える

という条件を満たすような方法を考える必要がある。

3.2 情報統合の原理

このような条件を満たす方法としては、各抽象度階層においてシステムが仮説を保持し、仮説どうしが条件付確率を介してネットワークを形成するような、確率的な情報統合モデルが適していると考えられる。そこで本稿では、各抽象度階層における情報や継時的な情報を統合して、対象についての最も尤もらしい内部像を推定するような処理を考える。このための基本的な問題は次のようなものである。

[基本問題] いま、行列 $P(a_i|b_j)$ および $P(b_j|c_k)$ で結ばれた三つの事象系 A, B, C があって、それぞれの事象系では事象 a_i $(i = 1,2,\cdots,L)$, b_j $(j = 1,2,\cdots,M)$, c_k $(k = 1,2,\cdots,N)$ が生起しているとする。ただし、B が定まったとき A と C とは独立であるとする。いま、観測者がこれら三つの事象系について観測を行ったところ、事象系 C について確率 $P_{t-2}(c_k)$ が、事象系 B について確率 $P_{t-1}(b_j)$ が、また事象系 A について確率 $P_t(a_i)$ が得られた。観測が終了した時点で最も尤もらしい a_i, b_j, c_k の組を求めよ (図 2)。

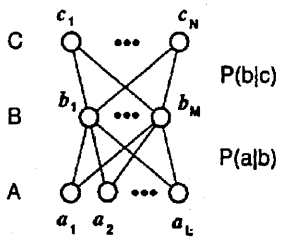

図 2: 情報統合の問題（例）

この基本問題を解くにあたっては、確率の双方向の伝搬を行う必要がある。例えば上の基本問題の場合、C の観測が得られた時点で a_i, b_j, c_k の状態は一意に定まるが、これに加えて B や A の観測が得られたとき、その観測結果によって C に対する推定結果も更新されるべきである。しかし、一般に、一つの確率ネットワークに対して双方向の確率の更新を安定かつ無矛盾に行うことはできない。そこで本稿では、確率伝搬の方法として、Pearl のベイジアンネットワークを用いる [5]。

いま、基本問題を拡張して図 3 のように木状に関係した事象系を考え、そのうち事象系 B に着目する。B の子孫の事象全体を D_B^-, それ以外の B 以外の事象を D_B^+ とすれば、事象 $B = (b_1, b_2, \cdots, b_M)$ の確信度 $BEL(B)$ は

$$BEL(B) = P(B|D_B^+, D_B^-) \tag{1}$$

と書ける。ここで、自然な仮定として、事象の独立性

$$P(D_B^+, D_B^-|b_j) = P(D_B^+|b_j)\, P(D_B^-|b_j) \tag{2}$$

を仮定すれば、ベイズの定理を用いて

$$P(B|D_B^+, D_B^-) = \alpha P(D_B^-|B)\, P(B|D_B^+) \tag{3}$$

と式変形することができる (本稿においては、ベクトルの積の表記は対応する要素どうしの積をとる演算を表すものとする)。ここで α は 形式的には $1/P(D_B^-)$ であるが、実際には結果が正規化されるように定めればよい。$\lambda(B) = P(D_B^-|B)$, $\pi(B) = P(B|D_B^+)$ とおけば、

$$BEL(B) = \alpha\, \lambda(B)\, \pi(B) \tag{4}$$

となる。

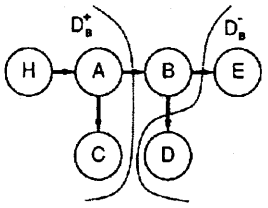

図 3: 木状に関係した事象系

後は、$\lambda(B)$ と $\pi(B)$ を求めればよい。まず $\lambda(B)$ について考える。B の k 番目の子をルートとする副木に含まれる事象を D^{k-} と書くと、

$$\lambda(B) = P(D_B^-|B) = \beta \prod_k P(D^{k-}|B) \tag{5}$$

となる (β は正規化定数)。但し、親の事象が定まった時の、子の間での事象の独立性を仮定している。ここで、いま仮に k 番目の子が事象系 E だったとすると、

$$P(D^{k-}|B) = \sum_i P(D_E^-|B, e_i)\, P(e_i|B) \tag{6}$$

$$= \sum_i P(D_E^-|e_i)\, P(e_i|B) \tag{7}$$

$$= \sum_i \lambda(e_i)\, P(e_i|B) \tag{8}$$

となるので、式 (5) と合わせると、親から子への条件付確率が与えられれば、漸化的に λ を伝搬できることが分かる。

次に $\pi(B)$ について考えると、

$$\pi(B) = P(B|D_B^+) \tag{9}$$

$$= \sum_i P(B|a_i, D_B^+)\, P(a_i|D_B^+) \tag{10}$$

$$= \sum_i P(B|a_i)\, P(a_i|D_B^+) \quad (11)$$

$$= \sum_i P(B|a_i) \left\{ \gamma\pi(a_i) \prod_m \lambda_m(a_i) \right\} \quad (12)$$

となる。ただし m は B を除く B の兄弟姉妹を数える添字であり、γ は正規化定数である。式 (12) の中括弧の中は、$BEL(A)$ の計算において必要なものであるから、$BEL(A)$ を計算した時点で分かっている。そこで、π についても、親から子への条件付確率が与えられれば、漸化的に π を伝搬できることが分かる。

結局、式 (4) において、親から子への条件付確率が与えられれば、確率としての性質に矛盾しない形で、双方向に確率を伝搬させられることが示された。仮定した条件は、ある事象系の事象が決まった時、その事象系の親と子の間の独立性 (式 (2))、および子どうしの間の独立性 (式 (5)) である。

3.3 仮説ネットワークの性質と特徴

上に述べた仮説ネットワークは、次のような性質を持っている。

1. 確率の伝搬が終了した平衡状態は、情報を与えた順序に依存しない。
2. 確率の伝搬は 1 回の拡散で終了するため、不安定な状態にはならない。
3. 計算量は、事象系の数に対しては線形、事象系内の事象数に対しては 2 乗のオーダーである。

このことから、先に述べた条件

1. 仮説を誰がどういう順序で作ったかには依存せず、
2. 従って処理の制御が容易であり、
3. 多くの緩い制約の積み重ねも効率的に扱える

が満たされていることが分かる。

常にいくつかの事象系が強い制約で押さえられる場合には、散層モデルにおける処理の制御のように、その事象系を駆動源にしてビームサーチのような仮説探索を行うことは効率的な制御方法であると言える。しかし、知覚的音源分離のように、状況によって強い制約となる事象系が異なったり、各々が緩い制約である場合などでは、このようなネットワークの方法が適すると考えられる。

3.4 仮説ネットワークの適用

まず、事象系を抽象度の階層に対応させ、図 4 のような構造を考える。階層は、下から C (Component) レベル、N (Note) レベル、および S (Scene) レベルである。S レベルは、リズムや和音など複数の単音にまたがる情報を表すが、これは 図 5 のようにいくつかの種類の異なる情報に分けてもよい。この場合構造が木ではなくなるが、このような単結合グラフに対しても上記の確率伝搬法が拡張できることが示されている [5]。これらの仮説ネットワークの構造を、本稿において仮説重合モデル (polymerized hypotheses model) と名づける。

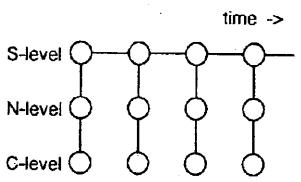

図 4: 仮説重合モデル (1)

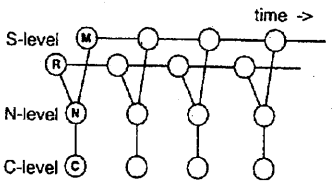

図 5: 仮説重合モデル (2)

各階層では、次のような仮説 (事象系における「事象」に相当) を保持する。

C レベルの仮説
"処理単位" における周波数成分データ。処理単位とは、入力音響信号を時間的に区分したものである。処理単位の境界は、一定時間 (例えば 200 ms) 以上周波数成分の立上り候補がない部分とする (それにまたがって単音が形成されることはないため)。仮説のバリエーションは、(1) 切断候補点 (4.1 節参照) で切断するか継続させるか、(2) 立上り時刻の評価、および (3) 周波数成分抽出のパワーの閾値などで生じる。

N レベルの仮説
処理単位における単音データ。単音形成クラスタリングにおけるパラメータ、音色同定パラメータによって仮説のバリエーションが生じる。

S レベルの仮説
処理単位ごとの音高情報 (和音情報) および時間情報 (リズム情報)。第 6 章参照。

また、仮説ネットワークに関係するモジュールとして、次のものが必要である。これらは、図 6 のように連係する。以下、着目する事象系を B とし、その親を A として説明する。

確信度ホルダ (B-Holder)
確信度 $BEL(B)$ を保持し、伝搬させる。

仮説クリエイタ (H-Creator)
ある B-Holder に対応する仮説 b_j を作り、初期確率 $\mu(b_j)$ を与える。一つの B-Holder あたり何個あってもよい。データが揃うなど、仮説生成が可能になった時点で起動する。

仮説コリレータ (H-Correlator)
隣接する B-Holder に関して、$P(b_j|a_i)$ を評価して行列 M を作る。隣接する B-Holder に対応する仮説が生成された時点で起動する。

仮説エバリュエータ (H-Evaluator)

ある B-Holder に対応する仮説の初期確信度を評価する。H-Creator は一つの B-Holder あたり何個あってもよいが、統一的な初期確信度を評価するために、H-Evaluator が一つの B-Holder あたり 1 個必要である。

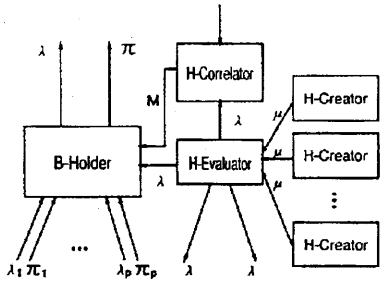

図 6: 仮説ネットワークの構成要素の連係

このうち、B-Holder の動作は次の通りである。B-Holder は、自分の親と子の有無と、(存在する場合には) それらとの通信のためのアドレスを認識している。親がいない B-Holder は、自分がルートであることを知っている。また、S レベルの B-Holder は、時間方向の子と階層方向の子とを区別することができるとする。

B-Holder は、H-Evaluator によって生成される。一度生成された後は、λ または π を受け取ることによって起動される。起動されたら、直ちに内部状態ベクトル $\lambda(B)$ および $\pi(B)$ を変更する。即ち

$$\lambda(B) = \beta \prod_{k=1}^{p} \lambda_k(B) \tag{13}$$

および

$$\pi(B) = \gamma M \pi_B(A) \tag{14}$$

とする。ここで M は $P(b_j|a_i)$ を要素とする行列であり、β, γ は正規化定数である。この結果、式 (4) によって $BEL(B)$ が更新される。次に、B-Holder は、隣接ノードに渡すべき $\lambda_B(A)$ および $\pi_k(B)$ を作成してこれを伝搬する。即ち

$$\lambda_B(A) = M^t \lambda(B) \tag{15}$$

および

$$\pi_k(B) = \zeta \pi(B) \prod_{j \neq k} \lambda_j(B) \tag{16}$$

とする。ここで M^t は M の転置を表し、ζ は正規化定数である。このとき、自分の起動の原因となったリンクは除き、他の全ての子と親にこれらを伝搬する。伝搬すべき相手が存在しなければ何もせず、再び λ または π によって起動されるまで休眠する。

ルートの B-Holder は、適宜時間方向の子を切り離し、自分の時間方向の子を新たなルートにすることができる。このための条件としては、時間方向の B-Holder の数に制限を設けるか、あるいは階層方向の子のエントロピー変化が僅少になった時点とすることが考えられる。時間方向の子を切り離した B-Holder は、その階層方向の子孫とともに消滅する。B-Holder が消滅した時点で、保持されていた仮説の状態 (確信度) は確定することになる。

4 ボトムアップ処理の概要

以上の章で、仮説ネットワークによる情報統合の原理を述べたが、本章、次章、および次々章において、仮説ネットワークに情報を与える処理モジュールの概要をごく簡単に述べる。

OPTIMA におけるボトムアップ処理の役割は、H-Creator としての仮説の生成と、H-Evaluator としての初期確信度の付与である。処理を分類すれば、

処理 1. 入力音響信号からの周波数成分の抽出

処理 2. 周波数成分の、重複を許したクラスタリング

処理 3. 処理 2 で生成したクラスタに対する、別途定めた特徴量に基づく類別 (教師なしクラスタリング) または判別 (教師ありクラスタリング)

処理 4. 処理 2 で生成したクラスタと処理 3 で行った類別や判別の結果に基づくシーン情報の抽出

の 4 種類となる。処理 2 において生成されるクラスタは個々の単音に対応しているので、これを単音クラスタ (Sound Cluster) と呼び、処理 2 の操作を単音形成クラスタリング (Clustering for Sound Formation) と呼ぶ。単音形成クラスタリングにおいて周波数成分の重複を許すのは、重複周波数成分 (shared component) を考慮するためである。また処理 3 において単音を類別または判別する操作は、楽器種類の同定に相当する操作であるので、これを音源同定クラスタリング (Clustering for Source Identification) と呼び、その結果生成されるクラスタを音源クラスタ (Source Cluster) と呼ぶ。以下、各処理について順に概観する。

4.1 周波数成分抽出

本稿においては、周波数成分は、スペクトログラムを時間–周波数–パワー空間における曲面とみなしたとき、その曲面上の一定の閾値を越えるピークについて、ピーク近傍における曲面の尾根方向の接線ベクトルを接続したものと考える。これを抽出するための操作として、以下に述べる挟平面回帰法を用いる。

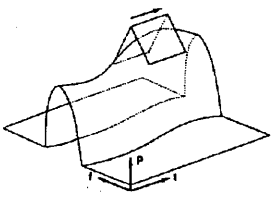

図 7: 挟平面回帰法

挟平面回帰法は、図 7 に示すように、スペクトログラム上のピークを 2 枚の平面で挟み込み、その 2 平

面の交線の方向ベクトルを利用し、次ピークの位置をあらかじめ予想することよってピーク接続の方向を定めるものである[6]。この挟み込みに使用する平面(これを挟平面と呼ぶ)は、最小2乗法により定めたピーク近傍の回帰平面である。この処理は、時間方向の局所的な先読みを許すことで処理効率を上げることができる。また、周波数成分を横軸時間、縦軸パワーとして見たとき、ある時点において

- パワー値が $m \cdot p_{max}$ 以下であること
- その前が立ち下がりであること
- その後 t_d 時間に、傾き g 以上の立ち上がりがあること

という条件が満たされる場合、その時点を切断候補点とし、切断した場合と継続させた場合の双方についての仮説を生成する。ここで m, t_d, g は定数であり、また p_{max} はその周波数成分のパワーの最大値である。

4.2 単音形成クラスタリング

本稿の範囲では、OPTIMAにおけるボトムアップ処理において、(1) 人間の聴覚的特性[7]および (2) 対象とする音の一般的な性質[8]に着目し、次の二つを入力に対し仮定する。

仮定 1. ひとつの単音に含まれる任意の周波数成分は、最も低い周波数成分に対して高調波関係にあること

仮定 2. ひとつの単音に含まれる全ての周波数成分の立上りが同時であること

ここで、高調波関係にない周波数成分を別の音とみなすことにすれば、衝撃音(打楽器音)などもともと高調波構造を持たない音や、基本波成分が欠けた音(missing fundamental)については対応できない。ボトムアップ処理においてはそのような音については特別な考慮をせず、トップダウン処理において対応するものとする。

さて、上の二つの仮定を置いたとき課題となるのは、高調波関係と立上り時刻の同時性という複数の特徴が周波数成分に存在したとき、これらを如何に評価してクラスタリングを行うかである。そこでわれわれは、単音形成クラスタリングにおける評価統合モデルを提案し、聴覚実験によって、基本的な場合におけるモデルの妥当性を示した[1]。

評価統合モデルは、図8に示すように、まず複数の特徴が独立に評価され、次にその評価値が統合されるとするモデルである。ここで、特徴の評価とは、ある周波数成分にその特徴が存在したときに分離知覚が生じる割合を評価することである。

評価統合モデルによって、任意の二つの周波数成分間に距離が定義される。これに基づいて行う単音形成クラスタリングのアルゴリズムとしては、いま考えている単音が高調波構造をなす(ある単音において、任意の周波数成分が、最も低い周波数の周波数成分に対して高調波関係にある)ことに注意すれば、次のような操作を導くことができる。

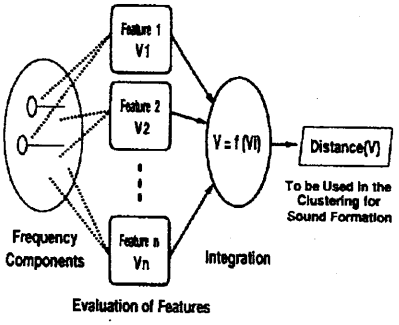

図 8: 単音形成クラスタリングの評価統合モデル

1. 最も低い周波数の周波数成分をクラスタ中心 C_1 とする
2. 周波数の低い順に周波数成分を走査し、C_1 との距離が θ_m より大きい周波数成分を見い出して、新たなクラスタ中心 C_2 とする
3. いずれのクラスタ中心に対しても、距離が θ_m より大きい周波数成分を見い出して、新たなクラスタ中心 C_3 とする
4. これを新たなクラスタ中心が見い出せなくなるまで繰り返す
5. 各クラスタ中心について、距離が θ_m を越えない周波数成分全てを見い出し、それぞれのクラスタに所属させる

ここで θ_m は別の音と判定するための確実度に対する閾値であり、0から1までの値をとる。0に近いほどいわゆる分析的な聞き方に近づく。また5番目の操作で、立上り時刻がクラスタ中心の周波数成分の立上り時刻よりも前にある成分については、立上り時刻に関する評価値を算入しないものとする。これは重複周波数成分を考慮するためである。以上のような操作により、人間がひとつの音と聞く可能性の高い周波数成分がクラスタ化され、ボトムアップに単音クラスタが形成される。

4.3 音源同定クラスタリング

音源同定クラスタリングとしては、教師の有無も含め、種々の方法を考えることができる。教師ありの場合は、予め蓄積しておいた楽器モデルを用い、特徴空間において判別分析を行って楽器種を判別する。教師なしの場合は、例えば特徴量に基づく凝集型の階層的クラスタリングによって単音クラスタを類別する。現在これらの両方の方法が実装されている。

4.4 シーン情報の抽出

シーン情報としては、単音の音高情報(和音情報)および単音の時間情報(リズム情報)の二種類を検討しているが、ボトムアップ処理で仮説を生成するのはこのうちの和音情報である。処理単位における単音の

情報から和音仮説を生成する処理 (即ち和音認識) を行う。

5 トップダウン処理の概要

OPTIMA におけるトップダウン処理の役割は、H-Creator としての仮説の生成と、H-Correlator としての条件付確率の評価である。

5.1 単音の予測

シーン情報、即ちリズム情報および和音情報に基づいて、単音の時刻および音高を予測して、仮説の生成および条件付確率の付与を行う。和音情報の利用に際しては、ある和音が定まった時の単音の出現確率を蓄積した知識源を参照する。

5.2 周波数成分の予測

周波数成分の予測には、知識源のうちの音記憶を用いる。音記憶としては、高調波比率やエンベロープなど音の特徴を抽出してパラメータを蓄積することも考えられるが、現在は、そのような抽象化を行わず、次のようなデータを音記憶として蓄積している。

$$\begin{aligned} T_k &= \{a_{ij}\}, \\ a_{ij} &= (p_{ij},\ f_{ij}). \end{aligned} \tag{17}$$

即ち、ある単音の記憶 T_k は、パワー値 p、周波数値 f を要素とする 2 次元ベクトルを要素とする行列である。行列の各行はその音に含まれる周波数成分に対応し、各列は時間のサンプル点に対応するものとする。これを図 9 に示す。

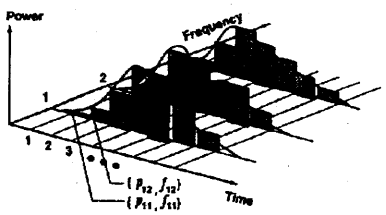

図 9: 音記憶

ここでは、蓄積された音記憶を用いて、仮説の生成および条件付確率の付与を行う。条件付確率は、生成された仮説と入力との間で予め定義された距離を測ることによって評価する。

6 時間方向の処理の概要

図 4 および図 5 に示すような仮説重合モデルでは、抽象度の階層方向の処理の他に、S レベルにおいて時間方向の処理を行う。

まず和音情報に関しては、知識源に蓄積された和音の遷移確率を参照して、和音仮説の生成および条件付確率の付与を行う。なお、和音の遷移において、ひとつの和音ごとの遷移を考えると一般に式 (2) の独立性が成り立たないと考えられるため、和音の遷移を N-gram で近似し (即ちある和音の出現確率は N 個前までの和音の状態にのみ依存すると近似し)、各ノードでは N-gram を状態として持つことにする。

次に時間情報に関しては、単音の時間情報の履歴を参照してリズム抽出を行い、単音時刻仮説の生成および条件付確率の付与を行う。リズム抽出の方法としては、Desain らの方法 [9] を参考にしたアルゴリズムを実装し、現在実験的検討を行っている。

7 処理モデルの動作

前章までに OPTIMA における情報統合の原理と、OPTIMA の各構成要素の概要とを述べたが、本章では、処理の進行に沿って処理モデルの動作を説明する。

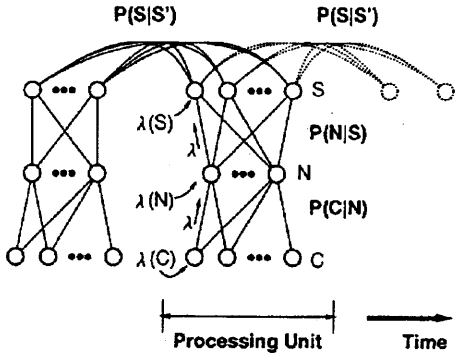

図 10: 処理モデルの動作

図 10 に、仮説重合モデルによる仮説ネットワークの一部分を示す (簡単のために 図 4 のような木構造とし、S レベルは和音情報に対応するものとする)。図 10 はある処理単位まで処理が進んだ時点を示す。図中のノード (○印) は B-Holder が保持する個々の仮説に対応し、ノード間のリンクは B-Holder 間に存在する λ リンクと π リンクの双方をまとめて表したものである。

ある処理単位における処理は、まず H-Creator が仮説を作ることにより開始する。H-Creator は、処理に必要なデータが揃うなど、起動可能な状態になった時に起動する。隣接する B-Holder に対応する仮説の生成が全て終了したことをトリガとして、トップダウンプロセスである H-Correlator が起動し、知識源を参照して $P(S|S'), P(N|S), P(C|N)$ を評価する。ここで $P(S|S')$ は和音の N-gram の遷移確率、$P(N|S)$ はある和音のときある音高の単音が出現する確率、また $P(C|N)$ はある単音のときある周波数成分の状態となる確信度である。

H-Correlator によって伝搬に必要な条件付確率が与えられると、ボトムアッププロセスである H-Evaluator が各仮説について確信度を評価し、B-Holder を生成するとともに確信度ベクトル λ を各階層に与える。例えば C レベルにおいては、周波数成分抽出プロセスから出力された確信度に基づいて $\lambda(C)$ が与えられる。確信度ベクトルが与えられると、B-Holder は、3.4 節に述べた手順に従って直ちにこれを伝搬する。次に、

N レベルと S レベルにおいて同様にボトムアップ処理の確信度ベクトル $\lambda(N), \lambda(S)$ が与えられ、$\lambda(C)$ と同様にこれらの情報が伝搬される。なお、仮説ネットワーク内で確信度ベクトルを与える順序は任意である。これを与えるごとに、ネットワーク内においてその時点で最も尤もらしい仮説のセットを求めることができる。

B-Holder に新たな情報 (確信度ベクトル) が与えられた時、情報の伝搬する様子を表したものが図 11 である。図 11 では、右から 2 番目の C レベルの B-Holder に情報が加えられている。情報は λ および π メッセージとしてネットワーク内に拡散し、一回の拡散で、確率としての整合性のある安定状態に達する。

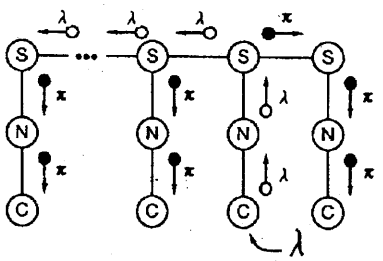

図 11: 確率の伝搬

本稿に示した処理モデル OPTIMA は、既に主要な処理モジュールの実装を終えており、現在情報統合の有効性を確認するための予備的な実験を行っている。例えば、音記憶 (Tone Memories) を用いた場合と用いない場合を比較した場合、これを用いた方がベンチマークテストにおける認識率が 10 % 以上改善されている。より詳細な評価実験については今後まとめて報告したい。

8 むすび

本稿では、知覚的音源分離の考え方に基づいて、ベイジアンネットワークによる情報の統合を基盤とする音響エネルギーの群化と構造化の機構を提案した。また、具体的応用例として、楽器音を対象とする単音記号列生成システムの処理モデル OPTIMA の全体像を示した。

知覚的音源分離に関連する従来の研究 [10, 11] では、複数種類の情報の統合という視点は考慮されていなかった。また、本稿に提案した処理モデルによるアプローチを黒板モデルに基づく類似の試み [12] と比較すると、まず、本稿の処理モデルでは処理の制御が極めて容易である点が特徴である。黒板モデルに基づくシステムがルールなどの形で処理の制御のための知識を必要とするのに対し、本稿の処理モデルは、システム全体を自律的に動くモジュールの集合として実装することが可能であり、グローバルな制御用の知識を必要としない。それに加え、本稿のモデルでは、得られた処理の結果に対して、各処理モジュールが出力する情報に基づく最尤推定という意味での定量的裏付けが与えらている点も特徴である。さらに、OPTIMA は、音源の位置[13] や音源の物理モデルなど利用可能な情報を必要に応じて逐次付加することが可能な枠組となっている。また、仮説ネットワークの構造も、単結合グラフの範囲内で拡張が可能である。

本稿では、処理モデル OPTIMA の要となる仮説ネットワークの原理と動作を中心に議論した。各処理モジュールの動作の詳細やシステム全体としての評価実験については、稿を改めて報告する予定である。

参考文献

[1] 柏野 邦夫, 田中 英彦: "2 つの周波数成分の分離知覚に関する工学的モデル～複数の要因の評価と統合～", 信学論 (A), **J77-A**, 5, pp.731-740 (1994).

[2] 柏野 邦夫, 田中 英彦: "モノラル楽器音の音源分離のための知覚的手がかりの検討と処理モデルの実装", 日本音響学会 聴覚研資 H-93-84 (1993).

[3] 中臺 一博, 柏野 邦夫, 田中 英彦: "音源分離システムにおけるパターン照合モジュールの並列実装と評価", 情処 94 春全大, 4T-6 (1994).

[4] 柏野 邦夫, 田中 英彦: "計算機への音楽の入力 - 「音」の分離抽出の難しさ-", 情報処理, **35**, 9, 印刷中 (1994).

[5] Pearl J. : "Fusion, Propagation, and Structuring in Belief Networks", *Artificial Intelligence*, **29**, 3, pp.241-288 (1986).

[6] 中臺 一博, 柏野 邦夫, 田中 英彦: "音楽音響信号を対象とする音源分離システム～音モデルに基づくアプローチ～", 情処研報 SIGMUS 1-1 (1993).

[7] Hartmann W. M.: "Pitch Perception and the Segregation and Integration of Auditory Entities", in Edelman, G. M. *et al (eds.)*: *"Auditory Function, Neurobiological Bases of Hearing"*, John Wiley & Sons, pp.623-645 (1988).

[8] 山口 公典, 安藤 繁雄: "短時間スペクトル分析法の自然楽器音への適用", 日本音響学会誌, **33**, 6, pp.291-300 (1977).

[9] Desain P. and Honing H.: "The Quantization of Musical Time: A Connectionist Approach", *Computer Music Journal*, **13**, 3, pp.56-66 (1989).

[10] Mellinger D. K.: *"Event Formation and Separation of Musical Sound"*, Ph.D. Thesis, Department of Music, Stanford University (1991).

[11] Brown G. J.: *"Computational Auditory Scene Analysis: A Representational Approach"*, Ph.D. Thesis, Department of Computer Science, University of Sheffield (1992).

[12] Nawab S. H. and Lesser V.: "Integrated Processing and Understanding Signals", in Oppenheim A. V. and Nawab S. H. (eds.) : *"Symbolic and Knowledge-Based Signal Processing"*, Prentice Hall, pp.251-285 (1992).

[13] Flanagan J. L., Johnston J. D., Zahn R. and Elko G. W.: "Computer-steered microphone arrays for sound transduction in large room", *J. Acoust. Soc. Am.* , **78**, 5, pp.1508-1516 (1985).