論　文

音楽情景分析の処理モデルOPTIMAにおける単音の認識

柏野　邦夫[†*]　　中臺　一博[†**]　　木下　智義[†]　　田中　英彦[†]

Note Recognition Mechanisms in the OPTIMA Processing Architecture for Music Scene Analysis

Kunio KASHINO[†*], Kazuhiro NAKADAI[†**], Tomoyoshi KINOSHITA[†], and Hidehiko TANAKA[†]

あらまし　音楽演奏の音響信号を対象として演奏情報を認識する試みとしては，従来自動採譜の研究が行われているが，複数種類の楽器音を含む音楽演奏を対象とする場合には，認識処理の有効性は極めて限られていた．そこで本論文では，複数種類の楽器音を含む音楽演奏の認識を音楽情景分析の問題としてとらえ，その解決を図る．ここで音楽情景分析とは，音楽演奏の音響信号から，単音や和音などの音楽演奏情報を記号表現として抽出することを指す．本論文ではまず，音楽情景分析を実現する上では情報統合の技術が不可欠であるとの認識から，ベイジアンネットワークによる情報統合の機構を備えた音楽情景分析の処理モデルOPTIMAを提案する．次に，特に単音の認識に的を絞って，提案する情報統合機構の有効性を示す．

キーワード　聴覚的情景分析，音源分離，情報統合，ベイジアンネットワーク，自動採譜

1. まえがき

本論文では，情報統合の機構を備えた音楽情景分析の処理モデルを提案し，特に単音の認識にかかわる処理に着目して，情報統合の有効性を示す．

一般に情景分析とは，感覚情報を入力として外界に生じている事象や外界に存在する物体に関する記述を出力する情報処理のことをいう．従来，情景分析の研究は主に画像情報を対象としていたが，近年，さまざまな音響情報の認識を情景分析の観点からとらえる考え方，すなわち聴覚的情景分析（auditory scene analysis）の枠組みが提案された[1]．聴覚的情景分析のうちで特に音楽音響信号を対象とするものを，本論文では音楽情景分析（music scene analysis）と呼ぶ．具体的には，音楽情景分析とは，音楽音響信号を入力とし，各楽器の演奏情報（単音，和音，リズムなど）を記号表現として出力する情報処理を指す．

計算機への音楽の演奏情報の入力に関しては，これまでにも自動採譜システムの研究が行われている[2]～[6]．しかし従来は，単一楽器の単旋律（ソロ歌唱など）か，または単一楽器の多重音（ピアノ演奏など）を対象とした研究が主であった．複数楽器の多重音を対象とする研究も試みられてはきたが[7]～[9]，音源の分離と同定が問題となるために，認識処理の有効性は限られていた．

そこで本論文では，複数楽器の多重音を含む音楽演奏を対象とする演奏情報の認識の問題を，聴覚的情景分析の観点からとらえて解決を図る．聴覚的情景分析において重要なのは，入力信号だけでなく，対象に関するモデルや統計的データなど利用可能なさまざまな情報を統合して総合的な判断を行うことである[10]．よって本論文では，情報統合の機構を示すこと，およびその機構の単音の認識に関する有効性を示すことの2点を目的とする．以下，まず**2.**において，提案する音楽情景分析の処理モデルの全体像を示し，**3.**において情報統合の機構を説明する．次に，**4.**と**5.**において，特に単音の認識に着目して，ボトムアップ処理の概要とトップダウン処理の概要をそれぞれ説明する．**6.**でシステムの動作例を示した後，**7.**において，単音の認識に関する評価実験を行い，ボトムアップ処理のみによる実験結果と，ボトムアップ処理とトップダウン処理とを併用した場合の実験結果とを比較することによって，提案する処理モデルの単音認識に対する有

† 東京大学工学部電気工学科，東京都
Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, 113 Japan
* 現在，NTT基礎研究所
** 現在，NTTソフトウエア本部

効性を示す．8. をむすびとする．

2. 処理モデル OPTIMA の全体像

2.1 構　　成

提案する音楽情景分析の処理モデル OPTIMA (Organized Processing toward Intelligent Music Scene Analysis) の全体像を図 1 に示す．OPTIMA は，各時点で得られた情報に基づいて，周波数成分 (frequency component)，単音 (musical note)，および和音 (chord) についての仮説を生成し，事後確率最大を評価基準として，全体として最ももっともらしい仮説の組を逐次求めていく枠組みである．図 1 に示すシステムの入力はモノラルの音楽音響信号であり，出力は，和音記号の列，楽器ごとに分類された単音記号の列，楽器ごとに分類された周波数成分の組，および拍位置を表す記号列である．これらの記号列は，音楽演奏に対する「知覚的な音」[10] に相当する．なお，出力される周波数成分をもとに，楽器ごとの音響信号波形を再合成することも可能である．

本処理モデルは，(A) 前処理部 (preprocesses)，(B) 主処理部 (main processes)，(C) 知識源 (knowledge sources)，および (D) 出力データ生成部 (output data generation) の四つの部からなる．

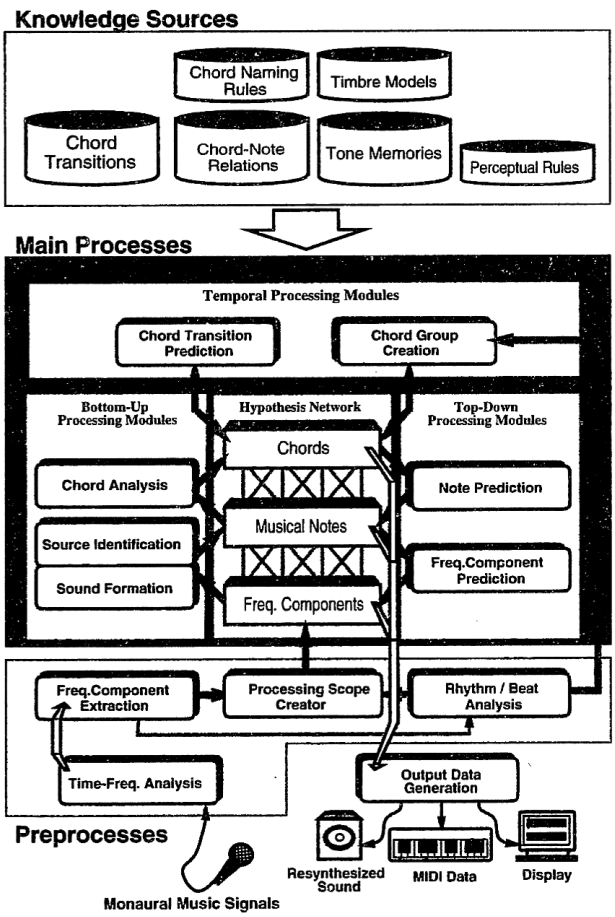


図 1　処理モデル OPTIMA の全体像
Fig. 1　The OPTIMA processing architecture.

前処理部は，入力音響信号を時間と周波数に関するエネルギー表現に変換すると共に，このエネルギー表現上における特徴を周波数成分として抽出し，リズム情報によりこれを整形して，主処理部に対する入力となる処理単位 (processing scope) を形成する部分である．ここで周波数成分とは，サウンドスペクトログラム上で，時間的に連続した，周波数方向に見たときのパワーの極大点の集合をいう．また処理単位とは，立上り時刻が互いに近接した周波数成分の集合を指す．

主処理部は，音響事象の仮説を保持するためのベイジアンネットワーク (仮説ネットワーク；hypothesis network) を備えている．仮説ネットワークは，(1) 周波数成分，(2) 単音，および (3) 和音の三つの抽象度の階層をもつ．単音は，個々の音符に対応する記号表現である．和音は，時間的に近接した複数の単音によって特徴づけられる記号表現である．仮説ネットワークに対して，(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 (bottom-up processing modules)，(b) 抽象度の高い階層から抽象度の低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 (top-down processing modules)，(c) 時間の推移に関する情報を扱う処理モジュール (temporal processing modules)，の三つの群に分けられる処理モジュールが情報を書き込む．ボトムアップ処理モジュールとしては，周波数成分の情報をもとに単音の情報を生成する処理 (単音仮説生成；sound formation および source identification)，単音の情報をもとに和音の情報を生成する処理 (和音仮説生成；chord recognition) の二つがある．トップダウン処理モジュールとしては，和音の情報をもとに単音仮説の確からしさに関する情報を出力する処理 (和音構成音情報付与；note prediction) と，単音の情報をもとに周波数成分仮説の確からしさに関する情報を出力する処理 (単音構成周波数成分情報付与；frequency component prediction) の二つがある．また，時間方向の処理モジュールとしては，和音の推移に関する情報を出力する処理 (和音遷移情報付与；chord transition prediction) と，時間的に連続する何個の処理単位が一つの和音を形成するかに関する情報を出力する処理 (chord group creation) の二つがある．これらの処理モジュールのうち本論文で扱うものは，単音仮説生成

と単音構成周波数成分情報付与の二つである．

主処理部における各処理モジュールは，それぞれ必要に応じて知識源を参照する．知識源としては，和音遷移に関する統計データ（和音遷移情報；chord transitions)，和音を構成する単音に関する統計データ（和音構成音情報；chord-note relations)，単音の集合に対しどのような和音名をつけるかをルールとしたもの（和音名ルール；chord naming rules)，単音を構成する周波数成分に関するデータ（単音記憶；tone memories)，音色を表現する特徴空間（音色モデル；timbre models)，および単音形成のための知覚的ルール（perceptual rules）を備える．これらのうち本論文で扱うものは，単音仮説生成モジュールの参照する知覚的ルールおよび音色モデル，単音構成周波数成分情報付与モジュールの参照する単音記憶である．

出力データ生成部は，主処理部の仮説ネットワークにおいて事後確率最大となった仮説を，画面表示やMIDI（Musical Instrument Digital Interface）データなど目的に応じた形で出力するためのものである．

2.2 他の聴覚的情景分析のモデルとの比較

聴覚的情景分析に関する研究[10]のうちで，ボトムアップ処理だけではなく種々の情報の統合を考慮したものとしては，中谷らによる音響ストリーム分離の試みと，Lesser らによる IPUS（Integrated Processing and Understanding of Signals）プロジェクトが挙げられる．

中谷らの実験システムは，単純な機能をもつ複数のエージェントから構成されるマルチエージェントシステムである[13],[14]．複数の処理モジュールの協調によって処理を行うという点では，中谷らのモデルと本モデルとは共通している．但し，中谷らのモデルが，自由度が高い反面で数量的定式化の困難な情報処理アーキテクチャを背景としているのに対し，本モデルは，情報統合の方法としてベイズの定理に基づく確率モデルを用いている．動作の理解と解析の容易さという点では，本モデルの方が見通しが良い．また，中谷らのモデルの出力が音源（話者）ごとに分離した音響信号であるのに対し，本論文のモデルでは，音響信号をもとに記号表現を生成して出力する点が異なっている．

また Lesser らの IPUS システムは，黒板モデルに基づいている[15]．本モデルも，複数のモジュールが共通の空間に対して情報の読出しおよび書込みを行うことで処理が進行するという点においては，黒板モデルと同様である．しかし Lesser らの黒板モデルでは，情報の統合および処理の制御は知識源として備えられた制御ルールによって行われており，情報統合のための定量的指針がない．このため，制御の安定性や処理の有効性を確保するための制御ルールの調整や保守が容易ではない．これに対し，本モデルでは，各処理モジュールは局所的な起動条件がそろったときに起動するだけであり，グローバルな制御ルールは不要である．また処理の保守としては，処理モジュールの独立性が確保されていることから，それぞれの処理モジュールにおける精度向上のみを考えればよい．

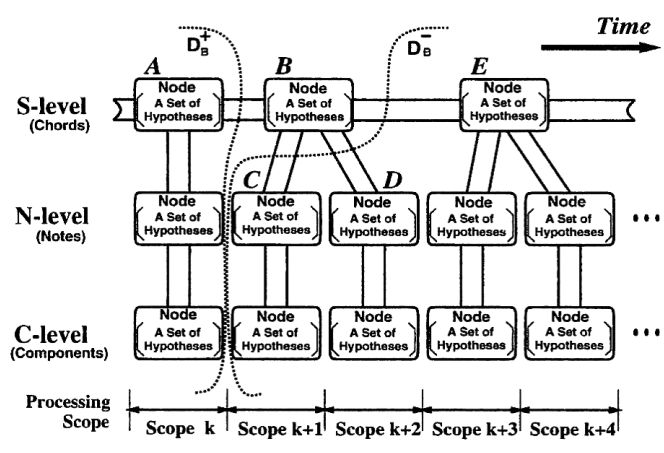
3. 情報統合のモデル

本章では，提案する処理モデルの主処理部に備わっている仮説ネットワークにおける情報統合の原理と，これに基づく処理モジュールの動作について説明する．

3.1 情報統合の原理

まず，処理の対象とする音楽演奏における抽象度の階層と時間的なつながりを考慮して，図2のような構造を考える．階層は，下から周波数成分（C）レベル，単音（N）レベル，および和音（S）レベルである．周波数成分レベルのノードと単音レベルのノードは1対1に対応するが，時間方向の複数の単音の並びが一つの和音を成し得るので，一般には単音レベルの複数のノードから和音レベルの一つのノードに対しリンクを設ける．

各ノードは，一般に複数の仮説を保持する．すなわち，周波数成分レベルのノードでは処理単位における周波数成分仮説，単音レベルのノードでは処理単位における単音仮説，和音レベルのノードでは和音のN-gram の仮説をそれぞれ保持する．これらのうち周



A,B,C,... は説明のためのラベルである（本文参照）．

図2 仮説ネットワークの構造
Fig. 2 Structure of the hypothesis network.

波数成分仮説は，処理単位における周波数成分の状態を個々の仮説としたものである．また単音仮説は，例えば「フルートの音高番号 60 とピアノの音高番号 48」といったように，各時点で同時に発音している単音の音源名と高さとを組にしたものである．

一方，各リンクは，隣接するノードにおける仮説同士の相関を，条件付確率として保持する．

以下 **3.1** の範囲の内容は，Pearl [16] に基づくものであるが，**3.2** で我々が提案する処理モジュールの構成と動作を理解する上で重要なので，要点のみを簡潔に記す．今，図 2 に示したノード B に着目する．B の子孫のノードに保持される仮説全体を D_B^- とし，B でも B の子孫でもないノードに含まれる仮説全体を D_B^+ とすれば，B に保持される仮説 $b=(b_1,b_2,\cdots,b_m)$ の確信度ベクトル $BEL(b)$ は，周囲のノードの仮説の状態が与えられた条件下での着目するノードの仮説の確率という意味で，

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

と書ける．なお本論文では，確信度という語を，式 (1) のように，各ノードにおいて保持される動的な条件付確率を指す場合に用いる．ここで

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

を仮定すれば（つまり，ノード B の仮説の状態が決まれば，B の親（子）側のノードの仮説の確率は B の子（親）側のノードの仮説の状態にかかわらず決まることを仮定すれば），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（本章においては，ベクトルの積の表記は，ベクトルの対応する要素同士の積を要素とするベクトルを得る操作を表すものとする）．ここで α は正規化定数である．$\lambda(b)=P(D_B^-|\,b)$，$\pi(b)=P(\,b\,|D_B^+)$ とおけば，

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

となる．$\lambda(b)$ はノード B とその子側のノードとの関連を，また $\pi(b)$ はノード B とその親側のノードとの関連を表している．

式 (4) より，$BEL(b)$ を求めるためには，$\lambda(b)$ と $\pi(b)$ を求めればよい．そこでまず $\lambda(b)$ 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仮説を D^{k-} と書くと，

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

となる（β は正規化定数）．但し，親の仮説が定まったときの，子の間での仮説の独立性を仮定している．ここで，今仮に k 番目の子がノード E だったとすると，

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

となることが示せるので，式 (5) と合わせると，親から子への条件付確率（すなわち $P(e_i|b_j)$ など）が与えられれば，漸化的に λ を伝搬できることがわかる．

次に $\pi(b)$ について考えると，

$$\pi(b)=\sum_i P(\,b\,|a_i)\left\{\gamma\pi(a_i)\prod_m \lambda_m(a_i)\right\} \tag{7}$$

となることが示される．但し 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 (7) 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，$BEL(a)$ を計算した時点でわかっている．そこで，π についても，親から子への条件付確率（すなわち $P(b_j|a_i)$ など）が与えられれば，漸化的に π を伝搬できることがわかる．

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することによって確信度ベクトルが求まることがわかった．仮定した条件は，あるノードの仮説の状態が決まったとき，そのノードの親と子の間の独立性（式 (2)），およびそのノードの子同士の間の独立性（式 (5)）である．本論文の場合，和音レベルにおいて和音の N-gram 仮説を保持しており，また一般に単音が決まれば和音は周波数成分の状態にかかわらず定まると考えることができるので，前者の独立性の仮定は妥当なものである．一方，後者の独立性を仮定していることから，単音の時間的なつながりは考慮されていない．

音楽情景分析システムにおける情報統合の機構は，仮説生成の順序に自由度があり，処理の制御が容易であり，かつ数多くの緩い制約を効率的に扱えることが必要である．Pearl のベイジアンネットワークは，このような条件を満たしているので，音楽情景分析に適した手法である．

3.2 処理モジュールの動作

以上の原理に沿った計算を行うために，仮説ネット

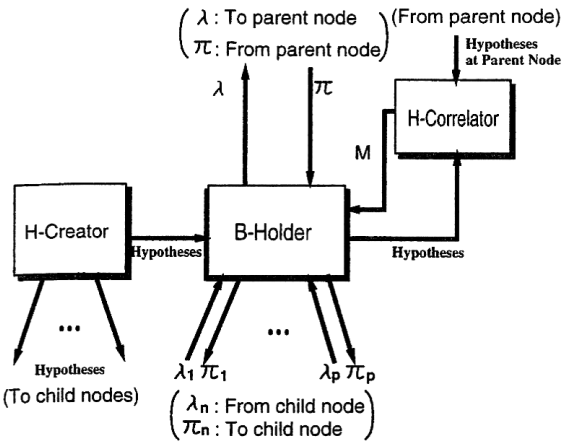


図3　1ノード当りの仮説ネットワークの構成要素
Fig.3　Modules required for a node of the hypothesis network.

ワークの各ノードにおいて次のようなモジュールを設ける．これらは，図3のように連係する．以下，着目するノードを B とし，その親を A として説明する．

（1）　確信度ホルダ（B-Holder）：　確信度ベクトル $BEL(b)$ を保持し，伝搬させる．仮説ネットワークの一つのノードに一つずつ存在する．

（2）　仮説クリエータ（H-Creator）：　ある確信度ホルダに対応する仮説 b_j を作り，確信度の初期値を与える．

（3）　仮説コリレータ（H-Correlator）：　隣接する確信度ホルダに関して，$P(b_j|a_i)$ を評価する．

本論文の例では，仮説クリエータとしてボトムアップ処理モジュールを用い，仮説コリレータとしてトップダウン処理モジュールを用いている．確信度ホルダは仮説ネットワークの実体である．

このうち，確信度ホルダの動作は次のとおりである．確信度ホルダは，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．親がいない確信度ホルダは，自分がルートであることを知っている．また，和音レベルの確信度ホルダは，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．

ある処理単位における処理は，まず仮説クリエータが確信度ホルダを生成し，これに仮説を与えることにより開始する．仮説クリエータは，処理に必要なデータがそろうなど，起動可能な状態になったときに起動する．一度生成された確信度ホルダは，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，式(5)および式(7)に従って内部状態ベクトル λ または π を変更する．この結果，式(4)によって $BEL(b)$ が更新される．次に確信度ホルダは，隣接ノードの確信度ホルダに渡すべき λ と π，つまり $\lambda_B(a)$ および $\pi_k(b)$ を作成してこれを伝搬する．すなわち

$$\lambda_B(a) = M^t\lambda(b) \tag{8}$$

および

$$\pi_k(b) = \zeta\pi(b)\prod_{j\neq k}\lambda_j(b) \tag{9}$$

とする．ここで ζ は正規化定数であり，M^t は $P(b_j|a_i)$ を要素とする行列 M の転置行列を表す．M は，仮説コリレータから与えられる．すなわち，隣接するノードの確信度ホルダに対する仮説の生成が終了すると，トップダウンプロセスである仮説コリレータが起動し，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率である．

このようにして，自分の起動の原因となったリンクは除き，他のすべての子と親に対して λ と π の伝搬が行われる．伝搬すべき相手が存在しなければ確信度ホルダは単に自らが保持している確信度ベクトルの変更のみを行い，再び隣接ノードからの λ または π によって起動されるまで休眠する．

ルートの確信度ホルダは，時間方向につながったノードの数が一定値に達した時点で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．時間方向の子を切り離した確信度ホルダは，その階層方向の子孫とともに消滅する．確信度ホルダが消滅した時点で，保持されていた仮説の確信度は確定することになる．

4.　単音仮説の生成

単音レベルの仮説クリエータは，ボトムアップ処理によって単音仮説を生成する．単音仮説生成処理は，単音形成処理（単音を形成する周波数成分のクラスタリング）と，音源同定処理（各単音についての楽器種類の判別）とに分けられる[17]．以下，これらの処理について順に検討する．

4.1　単音形成処理

本論文の単音形成処理では，入力に対し次の二つを仮定して，処理単位ごとに行う．

[仮定 1] 一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ほぼ高調波関係にあること

[仮定 2] 一つの単音に含まれるすべての周波数成分の立上り（開始端点）の時刻がほぼ同時であること

この仮定は，人間の聴覚的特性および対象とする音の性質から見て妥当なものである．すなわち音響心理学の知見によれば，人間の音源分離知覚（単音の形成）に関して，周波数成分の高調波関係の有無と周波数成分の立上り時刻の同時性の有無が音源分離知覚に強い影響を与えることがわかっている．また，音楽音響学の知見によれば，人間がピッチを知覚するような楽器音では，高調波の非調和性はおおむね 3%程度以下と考えてよく，また擦弦楽器，管楽器，打弦楽器のいずれにおいても，周波数成分の立上り時刻のずれは数十 ms 程度以下であることがわかっている [18]．なお，一部の打楽器音などのようにピッチを有しない音については，本論文では扱わないものとする．

さて，上の二つの仮定の下での検討課題は，周波数成分に高調波関係のずれと立上り時刻のずれという複数の特徴が存在したとき，これらをいかに評価してクラスタリングを行うかである．本論文では，単音形成クラスタリングにおける評価統合モデルを用いる [17]．評価統合モデルは，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り，等振幅の二つの周波数成分だけが存在するという最も基本的な場合に関しては，聴覚実験結果と対応することが示されている [17]．文献 [17] によれば，周波数成分の高調波関係のずれによって分離知覚の生じる確実性を c_h，立上り時刻のずれによって分離知覚の生じる確実性を c_o としたとき，

$$m = 1 - (1 - c_h)(1 - c_o) \tag{10}$$

によって双方の特徴が存在したときの分離知覚の確実性 m が得られるが，ここでは，m を周波数成分間の距離として単音形成のためのクラスタリングに用いる．クラスタリングは，前述の仮定 1 に注意すれば，次のような操作によって行うことができる．

（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 m_θ より大きい周波数成分を見出して，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 m_θ より大きい周波数成分を見出して，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 m_θ を超えない周波数成分すべてを見出し，それぞれのクラスタに所属させる

ここで m_θ は別の音と知覚するための確実性に対するしきい値であり，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（shared component；複数の単音に属する周波数成分が重なったために一つの周波数成分として観測されたもの）を考慮するためである．以上のような操作により，人間が一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る．

しかし，このような操作だけでは，ある単音の基本周波数が他の単音の基本周波数の整数倍になっている場合（同一またはオクターブ差の音程）には，これらを別の単音としてクラスタ化することができない．そこで，基本周波数が整数比となるような単音が含まれる仮説をも生成するものとした．この際，無限に多くの仮説が生じないようにするため，同時に発音する単音の最大数に制限を設けた．例えば，最大同時発音数を 3 としたとき，音高番号 60 の単音に対して (60,60)，(60,60,60)，(60,72)，(60,84)，(60,72,84)，(60,60,72)，(60,72,72)，… などの単音仮説を生成する．

4.2 音源同定処理

音源同定処理は，音色空間における判別分析によって行う．音色空間は，音楽音響の分野の知見や楽器の構造等を考慮して選択した 41 のパラメータに関して主成分分析を行うことにより構成した．選択したパラメータの主なものを表 1 に示す．主成分分析の寄与率は 95%とした．また，音色空間の次数を n としたとき，各音色は音色空間上で n 次元の正規分布として表すことができると仮定し，音色ごとの音色重心および分散・共分散行列を音色モデル（timbre model）として知識源に蓄積した．このとき，i 番めのサンプルが音色カテゴリー A に属する確率 P_{Ai} を式 (11) で算出することができる．

$$P_{Ai} = \frac{1}{(2\pi)^{m/2}\sqrt{|S_A|}} \exp\left\{-\frac{1}{2}D_{Ai}^2\right\} \tag{11}$$

表1　音色空間の構成に用いた特徴量の例
Table 1　Examples of parameters used for principal component analysis of timbres.

周波数成分に関する特徴量	周波数成分のパワーの比 周波数成分の立上り時刻の差 観測される周波数成分の数
パワー包絡に関する特徴量	立上りの傾き 振幅変化の度合（振幅変調度） 振幅変化の周波数

但し D_{Ai}^2 はマハラノビスの汎距離，m は音色空間の次数，S_A は分散・共分散行列を示す．同様の操作を他の音色カテゴリーに対して行うことにより，単音仮説がそれぞれの音色カテゴリーに属する確率を算出することができる．この確率値に基づいて，単音仮説生成時の初期確信度を与えた．

5.　単音仮説に基づくトップダウン処理

本章では，単音レベルと周波数成分レベルとの間の仮説コリレータの動作，すなわち単音に基づく周波数成分情報の付与について述べる．

5.1　知　識　源

単音情報に基づいて，出現する周波数成分を予測し，ある単音の下での周波数成分仮説に対する条件付確率の付与を行う．処理に際しては，単音に属する周波数成分を記憶蓄積した知識源（単音記憶：tone memories）を参照する．

単音記憶には，具体的な音の記憶として，次のようなデータを蓄積する．

$$T_k = \{a_{ij}\},$$
$$a_{ij} = (p_{ij},\ f_{ij}). \qquad (12)$$

すなわち，k 番目の単音記憶 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 a_{ij} を要素とする行列である．但しパワー値は，その単音記憶中の周波数成分の最大パワー値によって正規化されており，また周波数値は，その単音記憶の基本周波数の時間平均値との比を表す．行列の各行（i）はその音に含まれる周波数成分に対応する．行の数は楽器種類によって4～20程度である．また各列（j）は時間のサンプル点に対応する．本論文では10 msごとに1サンプルとし，最大80のサンプル点をとった．単音記憶の概念図を図4に示す．このような単音記憶を，クラリネット，フルート，ピアノ，トランペット，バイオリンの5種類の自然楽器音について，音域別に蓄積した．

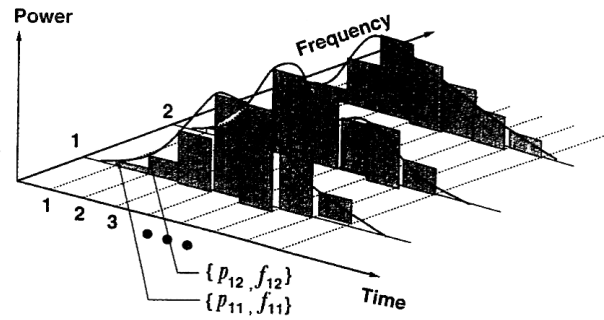


図4　単音記憶
Fig. 4　Tone memories.

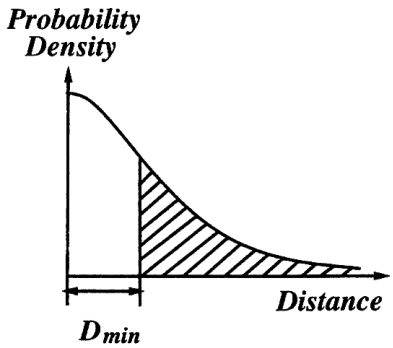
5.2　条件付確率の付与

単音仮説に基づいて周波数成分情報を付与する処理では，処理単位ごとに，あらかじめ蓄積した単音記憶を混合して，照合のための混合音（これを照合混合音という）を生成する．これは，式(12)に示した単音記憶のうち，単音仮説中に存在する可能性のあるものを選び，それらを実際のパワー値と周波数値に換算し混合することによって行う．このようにして生成したすべての照合混合音について，その処理単位における周波数成分仮説に対する距離 D を求める．混合に際しては，単音記憶同士の立上り時刻のずれや，単音記憶に含まれる各周波数成分の位相差も考慮した．すなわち，立上り時刻や周波数成分の位相をそれぞれ変化させたときの距離 D の最小値 D_{min} をもって，その照合混合音と周波数成分仮説との間の距離とみなした．ここで距離 D は，

$$D = \sum_{i=1}^{F}\sum_{j=1}^{N} \left|p'_{ij} - p_{ij}\right| \cdot f_{ij} \qquad (13)$$

と定義した．ここで，F は照合する周波数成分の数（照合混合音と処理単位の周波数成分仮説において周波数が対応する周波数成分は一つと数える），N は照合混合音の時間軸方向のサンプル点の数，p'_{ij} は処理単位の周波数成分のパワー，p_{ij}，f_{ij} は照合混合音の周波数成分のパワーおよび周波数である．この距離に基づいて条件付確率値を定める．

距離から確率への変換においては，混合音の周波数成分の分布に関し，距離尺度上での正規分布を仮定する．すなわち，混合音の周波数成分は，距離尺度上で，照合混合音の周波数成分を中心とし一定の分散をもつ正規分布をなすと仮定する（図5）．この場合，図5の網かけ部分の面積の2倍を，その照合混合音が与えられたときに，周波数成分が，照合した周波数成分仮説



混合音の周波数成分の分布に関し，距離尺度上で正規分布を仮定する．網かけ部分の面積の2倍を確率値とする．

図5　単音仮説の距離から確率値への変換
Fig. 5　Conversion of distance measure into probability.

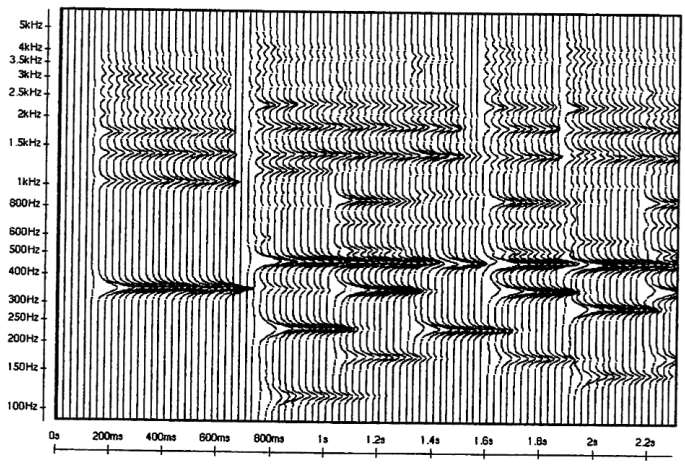


図6　周波数解析結果の例
Fig. 6　An example of spectrograms.

の状態となる確率（すなわち $P(C|N)$）とみなす．

6. システムの動作例

本章では，クラリネットとピアノによるアンサンブル演奏を入力した例を用いてシステムの動作を説明する．図6は，システムに入力された音響信号に対し周波数解析を行って得たスペクトログラムである．図7は，スペクトログラムに対して周波数成分を抽出した結果であり，周波数成分が線分として描かれている．図7中の縦の鎖線は，システムの前処理部で抽出した拍位置を示す．周波数成分のパワー値の変化と拍位置の情報を用いて，前処理部において処理単位が生成される．図7では，一例として入力開始から2番目の処理単位に属する周波数成分を太線で示している．処理単位が主処理部に入力されると，まず**4.**に述べた単音仮説クリエータが起動され，単音仮説生成される．単音仮説が生成されると，和音仮説クリエータと，**5.**に

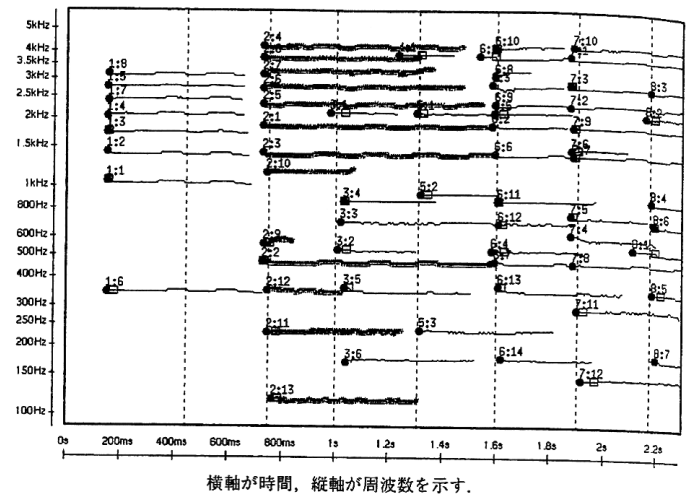


図7　抽出された周波数成分の例
Fig. 7　Frequency components extracted in the preprocessing block.

述べた単音レベルと周波数成分レベルとの間の仮説コリレータとが動作可能となるので，これらのモジュールが起動される．以下順次，図2に示したような仮説ネットワークが構成される．ネットワークが時間方向に一定の長さ（ノード数）に達すると，それより過去のノードが切り離され，仮説の状態が確定する．状態が確定した時点で，確信度最大の仮説がシステムから出力される．

7. 評価実験

本論文で提案する情報統合の機構に対し，単音レベルにおける情報統合の有効性を調べることを目的として，評価実験を行った．評価は，単音仮説生成モジュールだけを動作させた場合の単音認識結果（この場合，単音仮説のうちで初期確信度値が最も大きい仮説をシステムの認識結果とみなす）と，単音仮説生成モジュールと単音周波数成分情報付与モジュールの双方を動作させ情報統合を行った場合の単音認識結果とを比較することによって行う．

7.1 方　　法

評価用入力として，以下に述べる3種類の単音パターンを作り，これをサンプラで演奏した音響信号を作成した．ここでサンプラとは，任意の音響信号波形をそのままメモリに蓄積しておき，これを再生することによって発音する方式の音源装置である．本実験では，サンプラにクラリネット（cと表記），フルート（f），ピアノ（p），トランペット（t），およびバイオリン（v）の5種類の自然楽器音を蓄積した．単音パター

ンは，MIDI ノート番号 60 から 83 までの音域から一定数の音符（ここでは 2 音または 3 音）を選んで同時に発音する単音とし，これを時間方向に 50 個並べた．ここで同時とは，パターンを演奏する MIDI シーケンサ上で同じタイミングであることを意味する．各単音の継続時間は 750 ms とした．

単音パターンにおいては，各単音に由来する周波数成分の重なり方が単音認識の精度に大きく影響する．従って，ここでは単音パターンを次の三つの種類（クラス）に区別する．

（1） クラス 1 の単音パターン

同時に発音する単音の少なくとも 1 組が同一または整数倍の関係にある基本周波数をもつ（すなわちオクターブの関係にある）ような単音パターン

（2） クラス 2 の単音パターン

同時に発音する単音の少なくとも 1 組が 1.5 の整数倍の関係にある基本周波数をもつような単音パターンのうち，クラス 1 に属さないもの

（3） クラス 3 の単音パターン

クラス 1 にもクラス 2 にも属さない単音パターン

7.2 単音認識精度の指標

単音の認識精度の評価のための指標として，本論文では次に定義する正答指標 α，誤答指標 β，および認識率 R を用いる．

$$\alpha = \frac{a}{n}\,,\quad \beta = \frac{b}{n}\,,\quad R = \frac{1}{2}(\alpha - \beta) + \frac{1}{2} \qquad (14)$$

但し，n は入力（正解）に含まれる総音符数，a は出力に含まれる音符のうち音高と音色の両方が正しく認識された音符の数，b は出力に含まれる音符のうち，音高と音色のどちらかまたは両方が正しくない音符の数である．式 (14) における 1/2 の乗算と加算はスケール調整のためのものである．すなわち，システムが入力に含まれる音符と同数の音符を出力した場合，この正規化によって，すべてが誤っていれば $R = 0\%$，すべてが正しければ $R = 100\%$ となる．

7.3 結　　果

図 8，図 9，および図 10 に，単音仮説生成の精度を示す．各図において，グラフは棒 2 本で 1 組となっており，左側の棒はボトムアップの単音仮説生成処理のみを動作させた場合の結果を，また右側の棒はボトムアップの単音仮説生成処理に加えトップダウンの単音周波数成分情報付与モジュールを動作させて，単音レベルでの情報統合を行った場合の結果をそれぞれ示している．グラフの横軸の表記においては，最初の数字

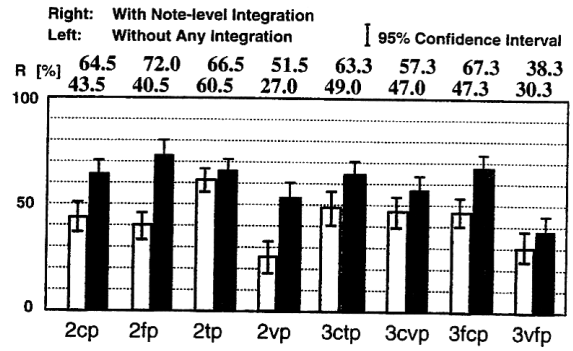


図 8　単音認識率の測定結果（クラス 1）
Fig. 8　Results of benchmark tests for note recognition (class 1).

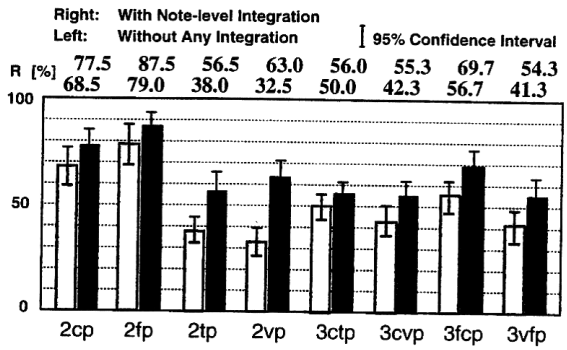


図 9　単音認識率の測定結果（クラス 2）
Fig. 9　Results of benchmark tests for note recognition (class 2).

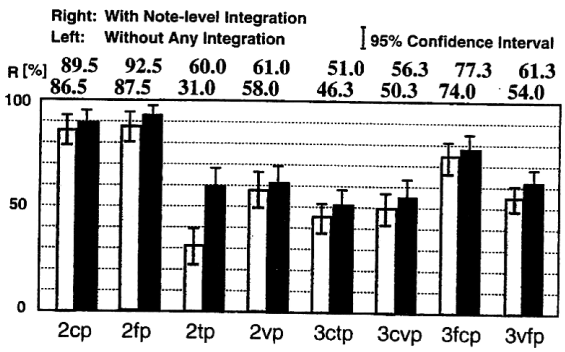


図 10　単音認識率の測定結果（クラス 3）
Fig. 10　Results of benchmark tests for note recognition (class 3).

が同時発音数，これに続くアルファベットが楽器音の種類を表す．例えば 2cp は，同時発音数 2 のパターンをクラリネットとピアノで演奏した場合の結果である．

図 11 に各クラスでの総音符数についての単音認識率を，また図 12 にクラス 1 の単音パターンに対する α と β をプロットしたグラフを示す．

認識率 R の値を見ると，各クラスとも，各グラフの左側に比較して右側の方が 3.0%から 31.5%（平均

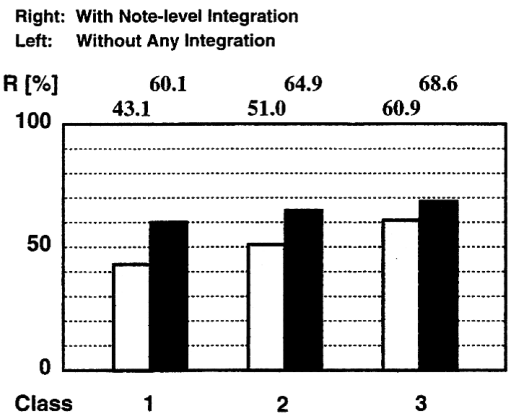


図 11 クラスごとの単音認識率
Fig. 11 Note recognition results for each class.

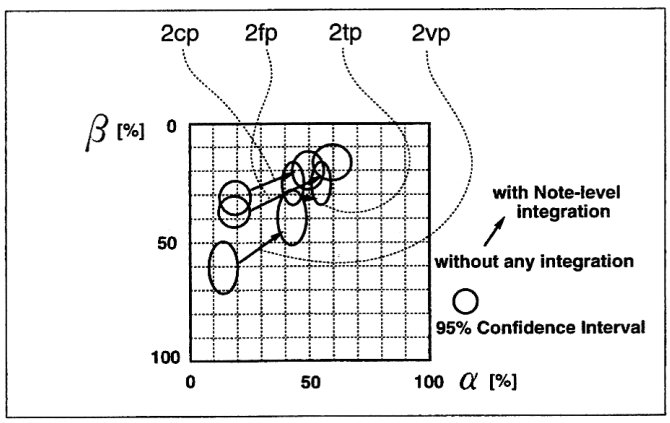


図 12 $\alpha - \beta$ 図（クラス 1）
Fig. 12 $\alpha - \beta$ plot (class 1).

12.9%）向上していることから，総じて単音レベルでの情報統合を行うことの効果が顕著に現れていると見ることができる．図 11 に示されるように，クラス別の認識率の向上の平均は，クラス 1 の場合 17.0%，クラス 2 の場合 13.9%，クラス 3 の場合 7.7%であり，周波数成分の重複する割合が大きくなるにつれて，単音記憶に基づく情報の統合の効果が大きくなっていることがわかる．

クラス 1（図 8）においては，少なくとも一つの単音の周波数成分は完全に他の単音の周波数成分と重複しているため，ボトムアップの単音仮説生成だけでは，ほとんど有効な処理結果を期待できない．実際，図 12 によれば，2cp，2fp，および 2vp において α 値が 20%以下にとどまっている．しかし，単音仮説生成部が基本周波数が整数倍の関係にある単音仮説を生成するため，単音構成周波数成分情報付与モジュールが有効に動作することができる．情報統合の結果，2cp，2fp，および 2vp では α 値，β 値ともに 20%程度，また 2tp でも 8%程度改善されている．

次にクラス 2（図 9）においては，同時に発音する単音において基本周波数成分は重複していないために，クラス 1 の場合に比べ単音認識率は一般に向上している．2fp，3fcp に対する結果に見るように，フルートの単音は比較的少数の周波数成分によって構成されるため，単音仮説生成において比較的高い認識率を得ている．

更にクラス 3（図 10）における実験結果を見ると，単音仮説生成の精度は，2cp，2fp，3fcp などのように，クラス 2 の場合よりも更に向上するものが多くなっている．

8. むすび

本論文では，音楽情景分析の処理モデル OPTIMA を提案し，ベイジアンネットワークによる情報統合の機構を示すと共に，単音の認識に最も関連の深い処理に絞って，情報統合の有効性を調べた．その結果，単音記憶の情報に基づくトップダウン処理を統合した場合，ボトムアップ処理のみの場合に比較して平均で 12.9%の単音認識率の向上が見られたことから，単音レベルにおける情報統合の有効性が示された．

しかし，本論文の実験で得られた単音認識率そのものは，いまだ実用的な値とは言えない．特に，トランペットのように基本周波数に対応する周波数成分のパワーが比較的低い楽器や，バイオリンのように周波数成分のパワーの著しい時間的変化や音色の揺らぎを有する楽器においては，クラリネットなど比較的定常な周波数成分を有する楽器に比べて単音認識精度が低い傾向が見られる．また，フルート演奏では息の音，クラリネット演奏ではピストンの操作音などが不可避的に混入するが，これらも単音認識精度を下げる一因となっている．また，ピアノのように，周波数成分が高調波関係であるという近似が，周波数が高くなるにつれて成り立たなくなる傾向にある楽器[19]においては，単音形成処理における誤りも生じている．今後，これらの問題への対策が必要である．

本論文で述べた処理モジュールは，図 1 に示した構成要素のうちの一部分である．我々は，主処理部における和音レベルの情報統合についても有効性を示す実験結果を得ているので，これについて稿を改めて報告する予定である．

文　献

[1] A.S. Bregman, "Auditory Scene Analysis," MIT Press,

1990.
[2] M. Piszczalski and B.A. Galler, "Automatic music transcription," Computer Music Journal, vol.1, no.4, pp.24–31, 1977.
[3] 新原高水，今井正和，井口征士，"歌唱の自動採譜，" 計測論，vol.20, no.10, pp.940–945, 1984.
[4] B. Mont-Reynaud, "Problem-solving strategies in a music transcription system," Proc. of IJCAI85, pp.916–918, 1985.
[5] C. Roads, "Research in music and artificial intelligence," ACM Computing Surveys, vol.17, no.2, pp.163–190, 1985.
[6] 片寄晴弘，井口征士，"知的採譜システム，" 人工知能学会誌，vol.5, no.1, pp.59–66, 1990.
[7] 長束哲郎，才脇直樹，井口征士，"異種楽器を対象とした採譜システム，" 信学 '92 春大会，D-499, 1992.
[8] 植田　護，橋本周司，"ブラインドデコンポジション問題としての音源の分離と同定，" 情処研報（93-MUS-3），vol.93, no.93, 1993.
[9] G.J. Brown and M. Cooke, "Perceptual grouping of musical sounds: A computational model," Journal of New Music Research, vol.23, pp.107–132, 1994.
[10] 柏野邦夫，"計算機による聴覚の情景分析 — はじめの一歩，" 日本音響学会誌，vol.50, no.12, pp.1023–1028, 1994.
[11] S. Handel, "Listening," MIT Press, 1989.
[12] W.M. Hartmann, "Pitch Perception and the Segregation and Integration of Auditory Entities," in G.M. Edelman, et al. (eds.), "Auditory Function, Neurobiological Bases of Hearing," pp.623–645, John Wiley & Sons, 1988.
[13] T. Nakatani, H.G. Okuno, and T. Kawabata, "Auditory stream segregation in auditory scene analysis with a multi-agent system," Proc. of 12th National Conf. on Artificial Intelligence, pp.100–107, 1994.
[14] 中谷智広，奥乃　博，川端　豪，"音環境理解のためのマルチエージェントによる調波構造ストリームの分離，" 人工知能学会誌，vol.10, no.2, pp.232–241, 1995.
[15] V. Lesser, S.H. Nawab, I. Gallastegi, and F. Klassner, "IPUS: An architecture for integrated signal processing and signal interpretation in complex environments," Proc. of 11th National Conf. on Artificial Intelligence, pp.249–255, 1993.
[16] J. Pearl, "Fusion, propagation, and structuring in belief networks," Artificial Intelligence, vol.29, no.3, pp.241–288, 1986.
[17] 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），vol.J77-A, no.5, pp.731–740, 1994.
[18] 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，vol.33, no.6, pp.291–300, 1977.
[19] 安藤由典，"新版楽器の音響学，" 音楽之友社，1996.

（平成 7 年 10 月 20 日受付，8 年 4 月 9 日再受付）

柏野　邦夫　（正員）

平 2 東大・工・電子卒．平 7 同大大学院電気工学専攻博士課程了．工博．同年 NTT に入社，基礎研究所情報科学研究部勤務，現在に至る．聴覚的情景分析の研究に従事．音響的情報を対象とする信号処理および知識処理に興味をもつ．平 6 情報処理学会奨励賞受賞．情報処理学会，人工知能学会，日本音響学会，IEEE 各会員．

中臺　一博　（正員）

平 5 東大・工・電気卒．平 7 同大大学院情報工学専攻修士課程了．同年 NTT に入社，ソフトウエア本部勤務，現在に至る．在学中，聴覚的情景分析の研究に従事．情報処理学会，人工知能学会，日本音響学会各会員．

木下　智義　（学生員）

平 7 東大・工・電子情報卒．現在同大大学院情報工学専攻修士課程在学中．聴覚的情景分析の研究に従事．情報処理学会会員．

田中　英彦　（正員）

昭 40 東大・工・電子卒．昭 45 同大大学院博士課程了．工博．同年東大・工・講師，昭 46 同大助教授，昭 62 同大教授，現在に至る．この間昭 53〜54 ニューヨーク市立大客員教授．計算機アーキテクチャ，並列推論マシン，帰納推論，オブジェクト指向計算システム，分散処理，CAD 等の研究に従事．著書「非ノイマンコンピュータ」，「情報通信システム」，共著書「計算機アーキテクチャ」，「VLSI コンピュータ I，II」，「ソフトウエア指向アーキテクチャ」．情報処理学会，人工知能学会，日本ソフトウエア科学会，IEEE，ACM 各会員．