人工知能学会誌 Vol. 13 No. 6 (Nov. 1998) 〔別刷〕

論 文

Technical Papers

単音連繋確率ネットワークに基づく音楽演奏の音源同定

Sound Source Identification for Ensemble Music Based on Music Stream Networks

柏野 邦夫* 村瀬 洋*
Kunio Kashino Hiroshi Murase

* NTT 基礎研究所
NTT Basic Research Laboratories, Atsugi, 243-0198 Japan.

1997年9月26日 受理

Keywords: sound source identification, Bayesian network, information integration, automatic music transcription, music scene analysis.

Summary

Sound source identification is an important but difficult problem in sound source separation. It is also a problem in the symbolization of music performances which include multiple simultaneous notes. As a solution to this problem, this paper presents a new method that can significantly improve the precision of sound source identification for music. Identification is here defined as the recognition of instrument names for each note included in an ensemble music monaural (or stereo) signal. The key idea of the proposed method is utilizing musical context. First we define the "music stream" that corresponds to a sequence of notes as a basic representation of musical context. We then describe the Bayesian method to introduce the contextual information to sound source identification. Experimental results show that the proposed method improves the accuracy of the source identification task for three-part ensemble music signals from an average of 67.8 % to 88.5 % .

1. まえがき

われわれは，複数種類の認識対象が混在する場合の音の認識の例題として，音楽のアンサンブル演奏に対する音源同定の研究を進めている．ここで音源同定とは，演奏されている個々の単音（楽譜上における音符に相当）が何の楽器の音であるかを認識することである．音楽演奏に対する音源同定は，自動採譜 [片寄 96] をはじめとする各種の応用において必要な処理である．

音源同定の問題に対するアプローチとしては，まず，判別分析など，音色の特徴に基づく方法が考えられる．実際，これまでにも，聴覚モデルと Kohonen の自己組織化ニューラルネットを組み合わせて，楽器名を判定する方法や [Cosi 94]，周波数成分の立上り時刻のずれと，高調波の性質で定義される量とで2次元空間を作り，その空間上で音色を判定する方法 [Brown 94] などが提案されている．しかし一般の音楽のように複数の音が同時に発音している場合，周波数成分の重複などによって，それぞれの単音の特徴を正確に抽出することは困難であり，音色の特徴のみに基づく方法の精度には限界がある．このため，対象とする単音のスペクトルパターンを予めシステムに蓄積しておき，そのパターン（あるいはパターンの混合物）と入力パターンとの照合によって音源同定を図る方法 [後藤 94, 片寄 91, 中臺 93] や，さらにその方法を判別分析と組み合わせて用いる方法 [柏野 96a] が提案されている．このような方法では，周波数成分の重複による音源同定精度への悪影響を軽減することができる．しかし，一般に自然楽器の演奏では，楽器の個体差や演奏の表情付けなどによって楽器の音色が大きく変動する．このため，照合に基づく方法の精度にも限界があった．

そこで，われわれは「適応型混合テンプレート」に基づく音源同定の処理モデル Ipanema を提案した[柏野 97, Kashino 97]．これは，入力音響信号に対して自乗誤差最小の解釈を与えるように蓄積テンプレートを変形させることによって，楽器の個体差や音色の変動を吸収する方法である．この方法によって，実演奏に対しても有効な音源同定処理の可能性が示されたが，短い時間内の音響信号の情報だけを利用するこの方法では，依然実用に耐える処理精度は得られていなかった．

ところで，われわれが音楽のアンサンブル演奏を聞く場合，比較的小編成の演奏であれば，たとえ音楽の専門家でなくとも，何と何の楽器による演奏であるかを言い当てることは，さほど困難でないことが多い．ところが，実演奏[*1]の一部分（例えば 0.3 秒間）だけを切り出して聞いた場合には，何と何の楽器の音が含まれているかを言い当てることは意外に難しいものである．この現象は，人間が日常，音楽を「音の流れ」として聞くことによって，楽器種類などの解釈を進めていることを示唆するものである．

本論文の基本的な発想は，[Kashino 97] の方法における音源同定精度を改善するために，音楽のもつ「音の流れ」の情報を統合しようというものである．

これまでにも，「音の流れ」については，音響ストリーム分離という観点からの研究が報告されている [Abe 96, 中谷 97]．これらの研究では，音の流れは，スペクトル特徴の時間的連続性あるいは時間的一貫性に基づいて定義されていた．これに対し，本論文では，スペクトルのレベルではなく，音楽としての音の繋がりの利用を考える．これは，音楽のように複雑なスペクトルをもつ信号を処理対象とする場合，スペクトル特徴の連続性や一貫性は利用が難しいからである．

以下，**2 章** で，処理モデル Ipanema に基づく音源同定システムの構成を概観した上で，**3 章** において「音の流れ」を表すグラフ（単音連繋確率ネットワーク; music stream probabilistic network. 以後単に単音連繋ネットワークまたは MSN: music stream network という）の構成について議論する．**4 章** では，単音連繋ネットワークを用いた音源種類確信度の更新について説明する．**5 章** で動作例を，また **6 章** で評価実験結果を示し，**7 章** をむすびとする．

*1 本論文において実演奏とは，サンプラなどの電子楽器による演奏ではなく，人間が自然楽器を用いて行った演奏を収録したもの意味する．

2. 処理モデル Ipanema による音源同定

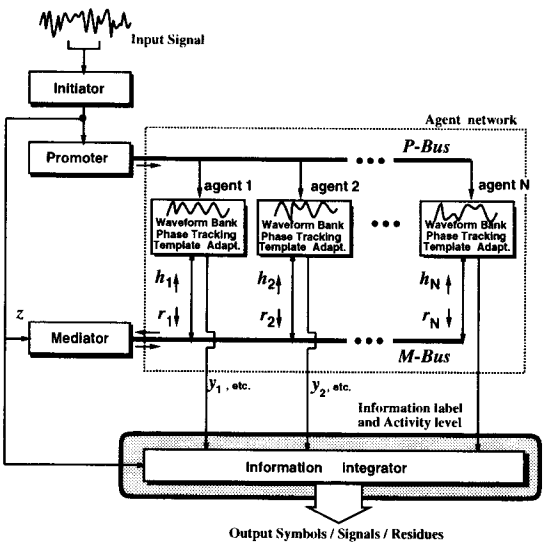
本章では，単音連繋ネットワークの位置付けを明らかにするために，音源同定の処理モデル Ipanema の全体像を示す．

2・1 システム構成

図 1 に，われわれが提案した処理モデル（アーキテクチャ）を示す [Kashino 97]．このシステムは，複数種類の音が混在した音楽音響信号を入力とする．出力は，楽譜に類似した形式の記号表現，各音源ごとの音響信号，および解釈の残差の音響信号である．

図 1 のアーキテクチャは，処理のきっかけを与えるイニシエータ（initiator），エージェントの処理を先導するプロモータ（promoter），音源分離・同定処理の主体となるエージェントネットワーク（agent network），およびエージェントの調停を行うメディエータ（mediator of agents）から成っている．そこで，図 1 のアーキテクチャを Ipanema と呼んでいる．また，上記の要素の他に，後処理モジュールとして情報インテグレータ（information integrator）が備わっている．情報インテグレータは，エージェントが出力する局所的な情報に加えて，音楽としての性質など，大域的な情報を統合して最終的な出力を生成するモジュールである．

本論文の主題は，情報インテグレータにおける単音



このうち本論文で対象とするのは，網かけの太枠で囲んだ Information integrator の部分である．

図 1 Ipanema アーキテクチャ

連繋ネットワークの作成と音源種類の確信度の更新である．

2・2 処理モジュール

〔1〕 イニシエータ

イニシエータは，入力信号を受け取り，音の立上りを検出する．立上りが検出されるごとに，入力信号波形を切り出して出力する．切り出された波形をフレームと呼ぶ．イニシエータによるフレームの生成は，後続の処理のきっかけとなる．

〔2〕 プロモータ

プロモータは，1フレームの波形を受け取り，周波数解析を行って，フレーム中に含まれている基本周波数成分を抽出する．フレーム中に複数の音が混在している場合には，基本周波数成分も複数存在する．プロモータは，抽出した基本周波数をプロモーションバス（P-Bus）に書き出す．この情報を P-Bus 情報と呼ぶ．P-Bus は，プロモータによって書き込まれ，次に述べるエージェントによって読み出される共通のデータ領域である．P-Bus 情報は，各エージェントが活動するかどうかを判断するために用いられる．

〔3〕 エージェント

Ipanema アーキテクチャでは，エージェントネットワーク中のエージェントは，個々の音源種類（例えばフルート，ピアノなど）に対応している．各エージェントはテンプレートバンクをもっており，テンプレートバンク中には，例えば半音ずつ基本周波数の異なる単音の波形が蓄積されている．

各エージェントは，随時 P-Bus を観察しており，プロモータによって書き出された P-Bus 情報を読み出す．P-Bus 情報中の基本周波数の値が，自分の担当する音源種類で発音可能な範囲内であれば，エージェントは担当音源が入力に含まれている可能性があるものと判断して活動状態となる．すなわち，テンプレートバンクから，基本周波数が現在の入力と最も近い波形を選び出し，位相トラッキング処理を行って位相同期テンプレート r_i を生成する．ここで，位相トラッキング処理とは，テンプレートバンク内の波形の位相を，入力中の対応する音源の位相に時々刻々合わせ込む処理であり，狭帯域のバンドパスフィルタを用いて実現している [Kashino 97]．一方，もし P-Bus 情報中の基本周波数が担当音源で発音不可能な範囲であれば，そのエージェントは何もせず，次の P-Bus 情報が準備されるまで休眠する．

活動状態のエージェントから生成された位相同期テンプレート r_i は，メディエーションバス（M-Bus）と呼ばれる共通のデータ領域に書き出される．M-Bus は，エージェントや次項に述べるメディエータによって読み書きされる共通のデータ領域である．エージェントが書き出した r_i はメディエータによって処理され，各エージェントに対応するフィルタ係数が求められるので，各エージェントは，そのフィルタ係数を M-Bus から読み込んで，r_i に対してフィルタ演算を行う．これによって 図2 に示したテンプレートフィルタリングが実現される．

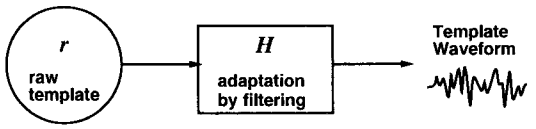


図2 音源波形のテンプレートフィルタリングモデル

エージェントからの最終的な出力は，テンプレートフィルタリングの出力波形 y_i，および記号表現のラベル（例えば「ピアノの C4」）である．

〔4〕 メディエータ

メディエータは，各エージェントの出力を調整する役割を負う．本論文においては，各エージェントの提案する位相同期テンプレートに対するフィルタ係数を返すことによって出力の調整が行われる．すなわちメディエータは，イニシエータから入力波形のフレーム z が切り出されてから一定時間待ち，その時間内に M-Bus に書き込まれた位相同期テンプレート r_i を読み込む．これらに基づいて，式 (1) の J を最小にするフィルタ係数を求める．

$$J = E\left[\left\{ z(k) - \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} h_n(m)\, r_n(k-m) \right\}^2\right] \tag{1}$$

ここで，k はサンプル時刻，$z(k)$ は入力信号波形，n は各エージェントに対応する添字，h は FIR フィルタの係数，M はフィルタの次数，N は活動状態のエージェントの数，E は時間平均を表す．J は，入力音響信号 z と，エージェントの出力の和との平均自乗誤差である．

この J が $h_n(m)$ に関して最小となるための必要条件は，全ての n と m に関して，偏微分 $\partial J/\partial h_n(m)$ が 0 となることである．この条件を用いると，$h_n(m)$ についての $N \times M$ 個の連立一次方程式

$$\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} E\left[r_i(k-j)\, r_n(k-m)\right] h_n(m) = E\left[r_i(k-m)\, z(k)\right] \tag{2}$$

を導くことができる (ここで, $i=\{0,1,\cdots,N-1\}$, $j=\{0,1,\cdots,M-1\}$ である).

メディエータは, 連立方程式 (2) を解くことによって, 各エージェントに対するフィルタ係数 h_i を算出し, これを M-Bus に書き込んでエージェントに返す.

〔5〕 **情報インテグレータ**

情報インテグレータは, エージェントネットワークの出力に対する後処理モジュールである. 情報インテグレータは, 各エージェントから, 波形 y_i および記号表現のラベルを受け取る. 基本的には, 入力フレーム波形とエージェント出力波形との相関値に基づいて, 最も相関値の高いエージェントのラベルを同定結果として出力することが考えられる. しかし, エージェントネットワークはフレームごとに独立に動作しているため, 単に最大相関のラベルを出力するだけでは, バイオリンのメロディーの流れの中で突然トランペットと誤認識された音が現れるなど, 音楽的に不自然な誤りを避けることができない. そこで, **次章**に述べるように, まず「音の流れ」を単音連繋として抽出して確率ネットワークを作り, 次にこのネットワークを利用して音源種類の確信度の更新を行う.

3. 単音連繋の抽出

単音連繋の抽出では, ある二つの単音の遷移が, 実際の旋律を分析した中で「どれだけありがちな遷移か」によって「音の流れやすさ」の評価尺度を定義する. すなわち, 二つの単音 n_{k-1}, n_k (ただし $k-1,k$ は発音開始時刻の順序を表す添字) が与えられたとき, この二つの単音の間に, 式 (3) によって定義される $Z(n_{k-1},n_k)$ を考える.

$$Z(n_{k-1},n_k)=W\sum_i\Big\{-w_i\log P_i(n_{k-1},n_k)\Big\} \tag{3}$$

ただし, i は Z において考慮する要因を数える添字であり, P_i は, 二つの単音の間が同じ旋律上に存在すると仮定したとき, 単音の遷移全体の中でその遷移が発生する確率を i 番目の要因について評価した値である. w_i (>0) は各要因に対する重みを表す. $-\log P_i$ は, 単音 n_{k-1} から n_k への遷移がもつ自己情報量を表すから, Z は, 自己情報量の重み付き線形和である. したがって, Z は, その二つの単音の遷移の現れにくさを表すと考えられる. そこで, 局所的に Z が最小となる方向に順次単音を繋いだものを, 単音連繋 (music stream) と定義する.

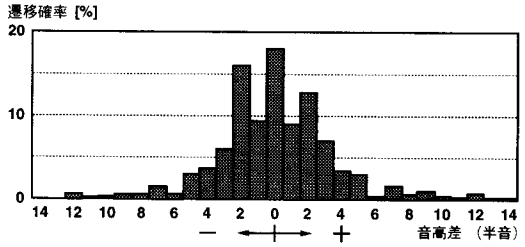


図3 音高の遷移確率

なお W は時間窓に相当し, 本論文では

$$W(\delta t)=\exp\left(\frac{\delta t}{\tau}\right) \tag{4}$$

と定義する. ここで δt は n_{k-1} の発音終了時刻と n_k の発音開始時刻との時間差の絶対値, τ は時定数である. 時間差が大きいほど W は大きくなる.

現在, 単音連繋形成の要因として, 以下に述べる (1) 音高遷移性, (2) 音色類似性, および (3) 役割同一性の三つを考慮している.

3・1 音高遷移性

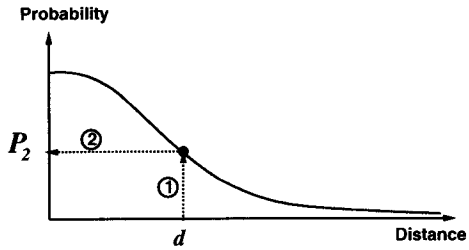
旋律中に現れる音高の遷移には, 現れやすい音程とそうでない音程とがある [松田 94]. そこで, 旋律中の音高の遷移確率を 式 (3) における P_1 として用いる. 本論文では, ポップス 196 曲, ジャズ 201 曲の計 397 曲の主旋律 (単音遷移数 62 689) を分析して遷移確率を得た. これを 図 3 に示す.

なお, このようにして蓄積された確率値は, 主旋律の音高遷移確率であって, ベースラインなど他の旋律や, 和音を演奏する楽器などには必ずしもあてはまらない. しかし本論文では, 簡単のため, 上記の確率値を全ての単音連繋についての P_1 として用いた. なお, 複数パートをもつ楽譜に対して統計をとったり, 別のジャンルの曲に対して統計をとったりし, 各パート別あるいはジャンル別に統計情報を適用すれば, 更に精度の高い遷移確率が計算できると考えられる.

3・2 音色類似性

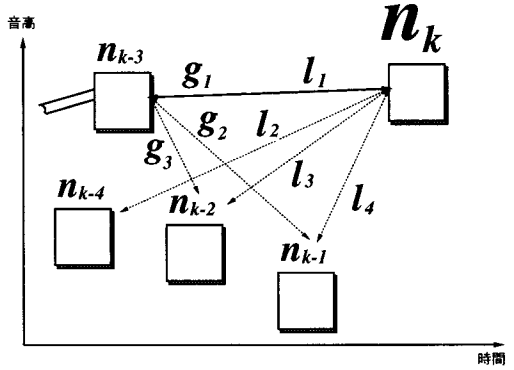
一つの旋律は, 類似した音色の系列によって作られることが多い. そこで, 単音の音色の間に距離を定義し, 同一の旋律中に, ある距離をもった二つの音の遷移が出現する確率を評価して, これを式 (3) における P_2 として用いる.

音色間の距離を定めるにあたって, i 番目のエージェントの出力波形と入力信号との相関値 v_i を要素とするベクトル V を考える. このベクトル V を音色ベクトルと呼ぶ. 音色ベクトルは, エージェントの総数 (活



単音連繋を形成する単音成分ベクトルの分布に関し，距離尺度上で正規分布を仮定し，距離 d から確率値 P_2 を得る．

図4 単音の距離から確率値への変換



新ノード n_k から既存ノードへのリンク候補（$l_1 \sim l_4$）のうちで，最小の Z 値を与えるものを選択する（l_1）．選択されたノード（n_{k-3}）からのリンク候補（$g_1 \sim g_3$）のうちで最小の Z 値を与えるリンク（g_1）が l_1 と一致するならば，そのリンクが単音連繋となる．既に n_{k-3} から g_1 以外の方向に単音連繋が生成されていれば，その連繋は切断され，$g_1(=l_1)$ に流れが変わる．

図5 単音連繋ネットワークの作成の説明図

動状態でないものも含む）だけの次元をもつ．音色間の距離は，単音 n_{k-1} の音色ベクトル V_{k-1} と単音 n_k の音色ベクトル V_k とのユークリッド距離と定義した．距離から確率への変換においては，単音連繋を形成する音色の分布に関して，距離尺度上での正規分布を仮定した．これを図4に示す．

3・3 役割同一性

アンサンブル演奏などの楽曲において，一連の旋律は，それぞれ主旋律，ベースラインといった，ある音楽的な役割を担っていると考えられる．そこで，一時点前まで，ある割合である役割を担っている単音連繋が与えられたとき，現時点の単音がその役割を果たす確率を評価して，式 (3) における P_3 として用いる．本論文では,「役割」として，同時に発音している単音の中での最高音となっているもの，および最低音となっているものを評価した．つまり，P_3 とは，n_k が最高（または最低）音である確率を表す．

P_3 は統計的分析によって得ることができるが，本論文の実験では，実験式として，

$$P_3 = aq + b \qquad (5)$$

とした．ここで，a,b は定数であり，q は，既存の単音連繋上の最大10個前までの単音に占める最高（最低）音であった単音の割合である．すなわち式 (5) は，一時点前まで最高（最低）音の割合が多かった単音連繋は，現時点でも最高（最低）音である可能性が高いことを近似的に表現したものである．

3・4 単音連繋ネットワークの形成

式 (3) を用いて，実際に単音連繋ネットワークを作成する手順を，図5を用いて説明する．図5は，新規にノード n_k が生成された時点を示している．新ノード n_k が生成されると，式 (3) を用いて，これと既存のノードとの間の Z が最小となるノードを選択する．選択されたノードから見ても新ノードとの Z が最小であれば，もし既存のリンクがあればそれを切断した上で，新ノードとの間にリンクを接続する．

このようにして，局所的に Z 値が最小となるノードどうしがリンクで接続され，単音連繋ネットワークが形成される．このネットワークは，音楽的には，おおむね各パートの旋律に対応すると考えることができる．

4. 単音連繋ネットワークによる音源確信度の更新

本章では，音の流れを考慮して音源同定を行う処理を，単音連繋上にある単音が次々に観測される各時点における，各単音の音源の事後確率を求める問題として考察する．

4・1 ベイジアンネットワークの枠組

ベイジアンネットワーク [Pearl 86] を用いると，関連のある一連の事象が観測された各時点における事象の事後確率を求めることができる．

時系列で与えられるデータに対する認識問題において，従来，動的計画法（DP）が盛んに用いられてきた．DP の定式化では，原理的に，経路の重み（通常，観測に基づく各仮説の確からしさと，仮説どうしの関連の高さとで決められる）が局所的に定義できなくてはならない．つまり，後になって観測された事実によって，既に観測されている仮説の確からしさ自体が変化

するような場合には，DP は適用することができない．これに対しベイジアンネットワークは，各仮説の確からしさを，観測が進むにつれて動的に変化し得る条件付確率として扱い，事象が次々に観測されていく各時点において，最もバランスよく制約を満たす仮説の組を求める手法である．しかも，模擬焼きなましなど繰り返し計算に基づく最適化手法に比較して高速（仮説数に対して多項式時間）かつ決定論的に計算を実行できる利点がある．以下，ベイジアンネットワークの原理の要点を述べる．

ノードとリンクからなる有向単結合グラフを考える．リンクの方向がノードの親子関係を表す．各ノードには，いくつかの仮説が対応しており，各リンクには，親のノードの各仮説が真であるという条件下での，子のノードの各仮説の確率を表す行列が対応している．例えば，親ノード A に 仮説 a_i $(i = 1, 2, \cdots, N)$ が対応していて，子ノード B に仮説 b_j $(j = 1, 2, \cdots, M)$ が対応しているとすれば，リンクは $P(b_j|a_i)$ を要素とする $M \times N$ 行列に対応する．

ここでは，各ノードの仮説における動的な条件付確率を確信度と呼び，リンクに与えられる静的な条件付確率と区別する．つまり，ノード A の確信度ベクトル $BEL(A)$ とは，A 以外のノードの仮説の状態（確信度ベクトル）が与えられた条件における，A の各仮説 a_i $(i = 1, 2, \cdots, N)$ に対する条件付確率を要素とする N 次元ベクトルのことをいう．このとき，Pearl のベイジアンネットワークのポイントは，$BEL(A)$ の各要素が，二つの値の積で決まり，それぞれが効率良く計算できるという点にある．すなわち，

$$BEL(A) = \alpha\, \lambda(A)\, \pi(A) \qquad (6)$$

である．ここで，α は正規化定数であり，$\lambda(A)$ と $\pi(A)$ は N 次元，つまりノード A の仮説数だけの次元をもつベクトルである．またベクトルの積の表記は対応する要素どうしの積を要素とするベクトルを求める演算を表すものとする．

式 (6) の $\lambda(A)$ および $\pi(A)$ は，実は，リンクで繋がれたノードをたどりながら順次求めることができる [柏野 96a, Pearl 86]．すなわち，$\lambda(A)$ は A の子のノードのλ ベクトルと条件付確率 $P(A$ の子のノードの仮説$|A$ の仮説$)$ を用いて計算でき，また $\pi(A)$ は A の親にあたるノードの π ベクトルと条件付確率 $P(A$ の仮説$|A$ の親のノードの仮説$)$ を用いて計算できる．これによって，各ノードでの確信度ベクトルを，親から子および子から親への 2 パスの情報の伝搬によって求めることができる．

4・2 単音連繋ネットワーク上の情報伝搬

本論文におけるベイジアンネットワークの具体的な適用法を 図 6 に示す．図 6 は，**前章**の方法によってリンク l が接続され，ノード n_k までの単音連繋が形成された時点を示している．

本論文では，ノードを 2 種類設ける．一つ目は，正方形で表されている確信度ノードである．確信度ノードは，前項で説明したベクトル λ とπ を保持する．二つ目は，台形で表されているデータノードである．データノードは，ネットワークに観測値を与えるために設けるノードであり，対応する確信度ノードの子として存在する．データノードでは，λ は観測値（ここでは音色ベクトル）に固定されており，π は等確率に固定されている．これは，観測値自体はネットワークの状態によらず不変だからである．また，対応する確信度ノード（親）との条件付確率も等確率となっている．すなわちデータノードは，確信度ノードに λ を与える役割のみをもつノードである．以下，単に「ノード」という場合には確信度ノードを意味する．

ノード n_k の λ ベクトルは，初期的にはデータノードのλ ベクトルである．リンク l ができると，まずノード n_{k-1} の π ベクトルを用いて，ノード n_k の π ベクトルが計算され，n_k に伝達される．次に，ノード n_k の λ ベクトルを用いて，ノード n_{k-1} の λ ベクトルが計算され，ノード n_{k-1} に伝達される．ノード n_{k-1} に伝達された λ は，更にその親側のノードに順次伝達されてゆく．伝達先がなくなった時点で，各ノードの λ ベクトルと π ベクトルが決まり，それらの要素どうしの積がその時点での音源確信度ベクトルとなる．

前節に述べたように，ベイジアンネットワークによる情報伝搬の枠組みでは，各リンクに対して条件付確率 P(子ノードの事象 | 親ノードの事象) を事前に与えることが必要である．この条件付確率は，統計的分析から収集することの可能なものであるが，本論文の実験では，これを次のように与えた．

$$P(h_j|h_i) = \begin{cases} \beta\left(\frac{1}{2} + c\right) & h_j \text{ と } h_i \text{ が同じ} \\ & \text{楽器名の場合} \\ \beta\left(\frac{1}{2} - c\right) & \text{上記以外} \end{cases} \qquad (7)$$

ここで，$P(h_j|h_i)$ は，親の仮説 h_i が真であったと仮定したとき，子の仮説が h_j となる確率，c $(0 \le c \le 1/2)$ は重み付け定数，β は正規化定数である．式 (7) は，単音連繋上にある単音が同じ楽器で演奏されるという仮説に対して，それらが異なる楽器で演奏されるという仮説よりも大きい条件付確率を与えることを表して

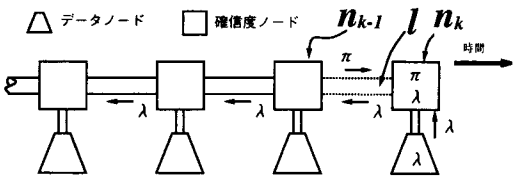


図6 確率の伝搬

いる．

5. システムの動作例

本章では，提案システムの動作例を示す．入力は，文献 [柏野 96b] に示される楽譜の「蛍の光」の 3 パートのアンサンブルを，バイオリン，フルート，ピアノの 3 種の楽器を用いて演奏したモノラルの音響信号である．図 7 は，情報インテグレータの稼働後の単音仮説の状態（曲の冒頭部分）を示す．各ノードの確信度ベクトルが棒グラフで表されている．図 7 では，各パートが単音連繋として適切に認識されていることが分かる．図 8 と 図 9 は，単音連繋ネットワーク使用前後の音源同定結果を楽譜に類似した形式で表示したものである．なお，音価の同定処理*2は行っていないので，すべて四分音符として実時間軸上に表示している．図 8 と 図 9 とを比較すると，単音連繋ネットワークの利用によって，いくつかの音符の音源同定誤りが正しく修正されていることが分かる．これは，単音連繋ネットワークにおける情報伝搬の結果，各音源の確信度が更新されたことによるものである．

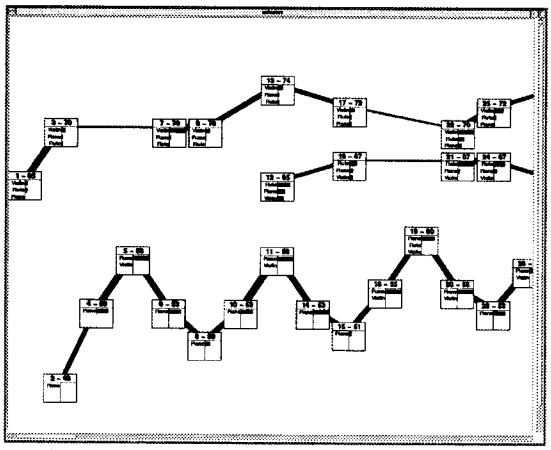
6. 評 価 実 験

表 1 のテスト曲を用いて，音源同定精度を調べた．これらの曲は，いずれも 3 パートのアンサンブルであり，各パートは単旋律となっている．例えば,「蛍の光」は [柏野 96b] に示される楽譜を演奏したものである．表 1 のうち「蛍の光」の収録では，編曲をプロの作曲家に，演奏を音大生と音大卒業生計 3 名に依頼した．また他の 3 曲の収録では，編曲を音大生に，演奏をプロの演奏家 3 名に依頼した．いずれの場合も，演奏に用いた楽器個体は，テンプレート用の単音を演奏した楽器個体とは別のものである．

本実験では，音源同定処理の精度を測るため，音高と時刻については人手で正解を与えて実験した（すなわち，イニシエータとプロモータは処理誤りを生じない状態に設定した）．実験に用いたパラメータ値を 表 2 に示す．これらの値は，予備実験の結果を参考にして人手で設定したものである．

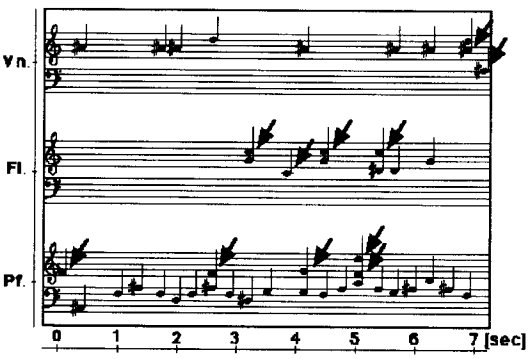
本実験では音高については正解を与えているので，出

*2 楽譜上の音の長さ，すなわち四分音符，八分音符などの音符の種類を判別する処理を，音価の同定処理と呼ぶ．



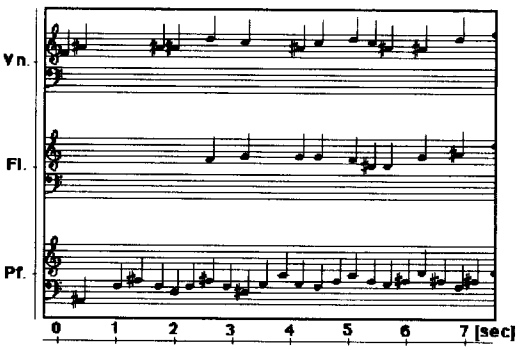
四角形が単音を表し，縦軸が音高，横軸が時間を示す．

図7 単音連繋ネットワーク使用後のノードの状態



矢印は，音源同定を誤っている音符を示す．

図8 単音連繋ネットワーク使用前の認識結果



表示範囲では，全ての音符が正しく音源同定されている．

図9 単音連繋ネットワーク使用後の認識結果

表 1　評価実験に用いたテスト曲

曲名	使用楽器（上のパートから順）	総単音数
アニー・ローリー *	Fl, Vn, Pf	234 音
ローレライ **	Fl, Vn, Pf	297 音
旅愁 ***	Vn, Fl, Pf	304 音
蛍の光 ****	Vn, Fl, Pf	242 音

Vn:バイオリン，Fl:フルート，Pf:ピアノ
* スコットランド民謡　　** Friedlich Silcher 作曲
*** J.P.Ordway 作曲　　**** スコットランド民謡

表 2　評価実験に用いたパラメータ値

$w_1 = 0.1$,	$w_2 = 1.1$,	$w_3 = 1.0$,
$a = 0.8$,	$b = 0.1$,	$c = 0.45$

力される単音数は入力に含まれる単音数に等しい（単音が余分に出力されたり欠落することがない）．そこで，認識率 R は，単に

$$R = \frac{(\text{音源名が正しく出力された単音数})}{(\text{出力された全単音数})} \tag{8}$$

とした．この定義は，出力される単音数が入力に含まれる単音数に等しく，かつ音高の誤りが生じない場合には，[柏野 96a] で用いられた認識率の定義と同じものである．

各単音の音源は，バイオリン，フルート，ピアノの 3 種類の楽器のうちのいずれかであることは既知とした．ただし各楽器の同時発音数については未知とした（すなわち，ある楽器が同時に複数音発音することも，まったく発音しないこともあり得るとした）．テンプレートフィルタリングのタップ数は 20 とした．

図 10 に実験結果を示す．これは，表 1 に挙げた各曲について式 (8) で与えられる認識率の値を算出した後，それらを平均した値を示したものである．これによれば，**3 章** で述べた三つの要因全てを用いた場合には，情報インテグレータを用いない場合に比べて，誤りを半分以下に減少させることができることが分かる．また，音色類似性については，単独で用いただけでは，音源同定精度に対してかえって足を引っ張る結果となっており，ネットワークの構造がパートと正しく対応していないことを反映していると考えられる．なお本実験では，単一の要因ごとに比較した場合，役割同一性の効果が他の二つの要因に比べて大きいが，これは，本実験に用いた演奏が役割同一性に適合していた（すなわち，曲の途中で主旋律やベースの楽器が交代しない）ことが理由として考えられる．

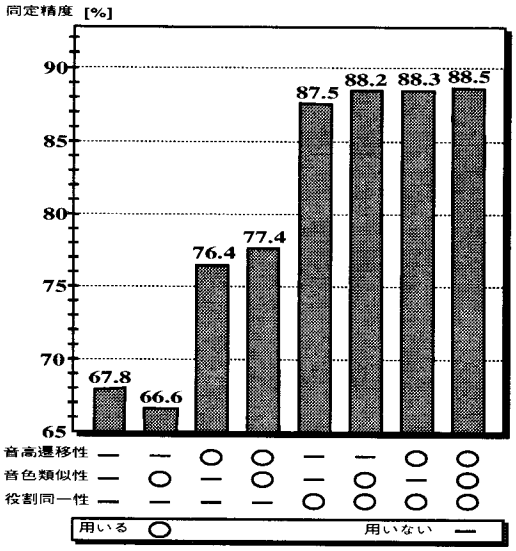


図 10　テスト曲に対する音源同定実験の結果

7. むすび

本論文では，音楽の実演奏に対する音源同定を目的として，音の流れ（単音連繋）の情報を統合することによって音源同定処理を高精度化する方法を提案した．アンサンブルの実演奏を用いた音楽認識テストの結果，単音連繋を考慮しない場合には平均 67.8 % であった音源同定精度が，88.5 % まで向上することが確かめられた．

しかし，自動採譜などの応用を考えれば，この音源同定精度はまだ十分な値とはいえない．今後，より実用に近い精度を実現するためには，本論文のように，単音の時間方向の局所的な繋がりを考慮するだけでは十分でないと考えられる．すなわち，同時に演奏されている音の関係を考慮することや，繰り返しパターンやフレーズなどといった音楽の構造に立ち入った処理を導入することも必要となるであろう．このような観点から，われわれは今後，単音の音高や音価の認識も含めて評価実験を行うとともに，更に精度の高い音源同定処理を目標として研究を進める予定である．

謝　辞

ご指導頂く NTT 基礎研究所の 東倉 洋一 所長および NTT 基礎研究所情報科学部の 石井 健一郎 部長，議論して頂いた NTT 基礎研究所の 奥乃 博 主幹研究員，川端 豪 主幹研究員 および NTT マルチメディアビジネス開発部の 中谷 智広 主査，音楽試料の収録

に協力頂いた 国立音楽大学の 兼 孝之 助教授および NTT 基礎研究所の 小坂 直敏 主幹研究員に感謝する.

◇参 考 文 献◇

[Abe 96] Abe M. and Ando S.: Application of Loudness/Pitch/Timbre Decomposition Operators to Auditory Scene Analysis, *Proc. ICASSP-96*, pp.2646–2649 (1996).

[Brown 94] Brown G. J. and Cooke M.: Perceptual Grouping of Musical Sounds: A Computational Model, *Journal of New Music Research*, Vol.23, No.2, pp.107–132 (1994).

[Cosi 94] Cosi P., Poli G. and Lauzzana G.: Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification, *Journal of New Music Research*, Vol.23, No.1, pp.71–98 (1994).

[後藤 94] 後藤真孝, 村岡洋一: 打楽器音を対象にした音源分離システム, 信学論 D-II, Vol.J77-DII, No.5, pp.901–911 (1994).

[柏野 96a] 柏野邦夫, 中臺一博, 木下智義, 田中英彦: 音楽情景分析の処理モデル OPTIMA における単音の認識, 信学論 D-II, Vol.J79-DII, No.11, pp.1751–1761 (1996).

[柏野 96b] 柏野邦夫, 木下智義, 中臺一博, 田中英彦: 音楽情景分析の処理モデル OPTIMA における和音の認識, 信学論 D-II, Vol.J79-DII, No.11, pp.1762–1770 (1996).

[柏野 97] 柏野邦夫, 村瀬 洋: 適応型混合テンプレートを用いた音源同定—複数楽器演奏への適用—, 信学技報, SP96-117 (1997).

[Kashino 97] Kashino K. and Murase H.: A Music Stream Segregation System Based on Adaptive Multi-Agents, *Proc. of Intl. Joint Conf. Artificial Intelligence*, Vol.2, pp.1126–1131 (1997).

[片寄 91] 片寄晴弘: 音楽感性情報処理に関する研究, 大阪大学基礎工学部 博士論文 (1991).

[片寄 96] 片寄晴弘: 自動採譜, 信学誌, Vol.79, No.3, pp.287–289 (1996).

[松田 94] 松田 稔, 秋山好一, 森 和義: 日本の楽曲の基本的特徴—音高について—, 音響誌, Vol.50, No.11, pp.897–905 (1994).

[中臺 93] 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).

[中谷 97] 中谷智広, 後藤真孝, 川端 豪, 奥乃 博: 残差駆動型アーキテクチャの提案と音響ストリーム分離への応用, 人工知能学会誌, Vol.12, No.1, pp.111–119 (1997).

[Pearl 86] Pearl J.: Fusion, Propagation, and Structuring in Belief Networks, *Artificial Intelligence*, Vol.29, No.3, pp.241–288 (1986).

〔担当委員：白井良明〕

著 者 紹 介

柏野 邦夫(正会員)

1990 年東京大学工学部電子工学科卒業. 1995 年同大大学院電気工学専攻博士課程修了. 博士(工学). 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. 音響認識, マルチメディア認識の研究に従事. 音響・画像情報を対象とする信号処理と知識処理に興味をもつ. 電子情報通信学会, 情報処理学会, 日本音響学会, IEEE 各会員. 〈kunio@ca-sun1.brL.ntt.co.jp〉

村瀬 洋

1978 年名古屋大学工学部電子工学科卒業. 1981 年同大大学院修士課程修了. 同年日本電信電話公社(現 NTT)入社. 以来, 文字・図形認識, コンピュータビジョン, マルチメディア認識の研究に従事. 1992 年から 1 年間米国コロンビア大客員研究員. 現在, NTT 基礎研究所情報科学研究部メディア情報認識グループリーダ. 工学博士. 1985 年電子情報通信学会学術奨励賞, 1992 年電気通信普及財団テレコムシステム技術賞, 1994 年 IEEE-CVPR 国際会議最優秀論文賞, 1995 年情報処理学会山下記念研究賞, 1996 年 IEEE-ICRA 国際会議最優秀ビデオ賞受賞. 電子情報通信学会, 情報処理学会, IEEE 各会員. 〈murase@apollo3.brL.ntt.co.jp〉