論　文

適応型混合テンプレートを用いた音源同定
——音楽演奏への適用——

柏野　邦夫†　　村瀬　洋†

Sound Source Identification by Adaptive Template-Mixture Method
——Application to Ensemble Music Recognition——

Kunio KASHINO† and Hiroshi MURASE†

あらまし　同時に複数の認識対象が混在する音の認識では，音源同定処理が必要である．本論文では，音楽演奏など，実環境における音の多様性や変動にも対処できる音源分離同定を行うことを目的として，適応型テンプレートを用いた音源同定処理を提案する．更に，この処理を応用して，同時に複数の音を認識対象とするシステムの代表例であるアンサンブル演奏の認識システムを構築する．構築したシステムに対し，自然楽器の和音によるベンチマークテストを行ったところ，単純なマッチトフィルタによる音源同定に比べ，音源同定精度の平均が55.0%から69.5%に改善された．またアンサンブルの実演奏を用いた音楽認識テストにおいても，提案手法の音源同定精度に対する改善効果が確認された．

キーワード　音源同定，音源分離，音楽情景分析，自動採譜，マッチトフィルタ

1. まえがき

音の認識の研究では，従来，認識の対象とする音の種類をあらかじめただ一つに限定するものがほとんどであった．例えば音声認識システムは，人間の音声だけを認識の対象とする．もちろん，音声認識システムの入力として，色々な雑音が混在していることを考慮することも多いが，その場合も，認識の対象となるのは音声だけである．

これに対し我々は，音楽演奏を例題として，複数種類の認識対象が混在する場合の音の認識に取り組んでいる．この問題は，シグナル（認識対象の音）とノイズ（認識対象ではない音）が一義的に決まっているのではなく，同時に複数の音がシグナルとなり得るのが特徴である．このような問題は，音楽の自動認識システムにとどまらず，音によって周囲の状況を理解するシステムや，マルチメディアデータベースの自動インデクシングシステムなどにおいても重要であると考えられる．

さて，複数種類の認識対象が混在する音の認識では，入力の音響信号から個々の音に相当する部分を分けて取り出すことと，個々の音が何の音であるかを判定することの二つの課題がある．本論文では，前者を音源分離（sound source separation），後者を音源同定（sound source identification）と呼ぶ．近年，音源分離の研究は盛んに行われており[1],[2]，また一方で，室内などの音響事象の認識[3]，話者認識，音声区間の切出しなど，単一で存在する音源の識別の研究も行われている．もし，混合音に対する完全な音源分離が実現され，原音源の信号波形が忠実に復元されるのであれば，音源同定の問題は，単一で存在する音源波形を対象とする場合のみを検討すれば十分である．しかし実際には，周波数成分の重複などにより，一般に完全な音源分離の実現は非常に困難である．

そこで我々は，必ずしも音源分離を前提とすることなく，他の音源の混在を考慮に入れて音源同定を行う方式を提案する．同時に複数の音源を認識する試みとしては，柏野らによるOPTIMAの研究がある[4],[5]．OPTIMAは，情報統合を鍵技術とする音楽認識の処理モデルである．これまでに，2〜3パートの編成のモノラルのアンサンブル演奏を入力とし，種々の情報を

† NTT基礎研究所，厚木市
NTT Basic Research Laboratories, Morinosato-Wakamiya, Atsugi-shi, 243-0198 Japan

統合してパートごとの音符情報などを出力する実験システムが実装されている．しかしながら，その評価実験は主にサンプラ(注1)の音を用いて行われていた．これは，実楽器音は多様で変動が大きいために，精度良く処理することが難しかったからである．

本論文では，多様で変動の大きい対象を扱うための鍵技術として新たに「適応」の考え方を導入する．以下 **2.** では，音源同定のためのテンプレートを入力に合わせて変化させるという適応型混合テンプレートのアイデアを提案し，問題の定式化を行う．**3.** では，計算を実行するための具体的なシステムの構成を議論する．**4.** では，構築したシステムに対し評価実験を行って，**2.** で提案する処理の効果を検討する．**5.** をむすびとする．

2. 適応型混合テンプレート

2.1 テンプレートフィルタリング

今，システムに入力される波形は，いくつかの音源波形の和であるとする．この入力波形を各音源波形に分離する問題において，各音源波形の典型的な例（テンプレート波形）$y_n(k)$ が与えられているとする．ここで n は各音源に対応する添字，k はサンプル時刻を表す．すると，問題は，

$$J = E\left[\left\{z(k) - \sum_{n=0}^{N-1} y_n(k)\right\}^2\right], \qquad (1)$$

の最小化として定式化することができる．ここで $z(k)$ は入力信号波形，N は音源の数，E は時間平均を表す．なお N はあらかじめ与えられてはいない．テンプレート波形 $y_n(k)$ のモデルとして，図 1 に示すような「テンプレートフィルタリングモデル」を考える．これは，テンプレート波形（template waveform）を，原テンプレート（raw template）とフィルタ H による変形とでモデル化するものである．フィルタとして FIR 型を用いることにすれば，

$$y_n(k) = \sum_{m=0}^{M-1} h_n(m)\, r_n(k-m), \qquad (2)$$

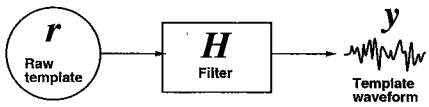
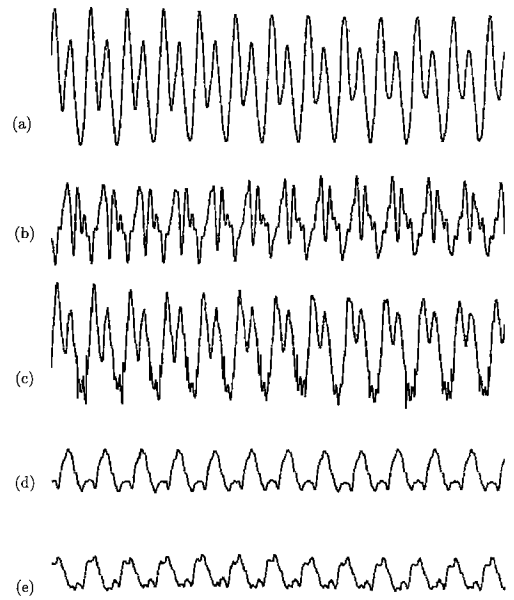


図 1　テンプレートフィルタリングモデル
Fig. 1　The template filtering model.

と書ける．ここで，h は FIR フィルタ H のインパルス応答，r は原テンプレート波形，M はフィルタの次数である．

一般に，処理対象の音源波形は多様であり変動するので，h や r として固定の値を用いることはできない．音源波形の多様性の例を図 2 に示す．もし位相を捨てて例えばパワースペクトル表現を用いることにしても，その表現の空間上で音が変動するという事情は基本的に同じである．従って，音源の変動に対処する何らかの仕組みが必要である．ここでは，フィルタの



(a) はヤマハ，(b) はベーゼンドルファのピアノである．どちらも同じ高さ（F4），同じ時間部分（立上りから 160～195 ms）であり，ほぼ同じ強さで弾いたものであるが，波形は異なっている．(a) の波形を式 (3) における z とし，(b) の波形，フルートの波形，およびバイオリンの波形を，式 (3) における $r_i\,(i=0,1,2)$ としてテンプレートフィルタリングを行った結果得られた $y_i\,(i=0,1,2)$ がそれぞれ (c)，(d)，および (e) である（この例では $N=3$，$M=160$）．(c) と (a) の相関値は，(d) と (a) の相関値や (e) と (a) との相関値に比べ高くなっている．この相関値を用いて音源同定が行える．サンプリング周波数は 48 kHz．

図 2　ピアノ波形の多様性とその吸収
Fig. 2　Examples showing the variability in piano waveforms and also the effects of template adaptation using template filtering.

(注 1)：サンプラとは電子楽器の一種であり，あらかじめ実楽器の単音波形を装置内に蓄積しておき，これを再生することによって発音する方式のものを言う．波形自体は実際の楽器のものであるが，基本的には同一条件では同一の波形が再生されるため，実際の楽器演奏に比べて音の波形の変動に乏しいという特徴がある．

係数 $h_n(m)$ を変えることを考える．式(1)を，式(2)を用いて書き直すと

$$J = E\left[\left\{z(k) - \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} h_n(m)\, r_n(k-m)\right\}^2\right] \tag{3}$$

となる．

この J が $h_n(m)$ に関して最小となるための必要条件は，すべての 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，$N \times M$ 個の連立1次方程式

$$\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} E\left[r_i(k-j)\, r_n(k-m)\right]\, h_n(m) = E\left[r_i(k-m)\, z(k)\right] \tag{4}$$

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）．この連立方程式は，係数行列の逆行列を求めることによって解くことができる．

2.2 位相トラッキング

前節の処理は，主として，原テンプレート r の基本周波数および位相が z に含まれている音源の基本周波数および位相と一致している場合において有効であると考えられる．なぜなら，前節のフィルタ H は，信号の周波数を変えることはできないからである．このため，原テンプレートの位相を，z 中の対応する音源の位相に時々刻々合わせ込むメカニズムが必要である．

もし，入力信号が，複数の音源からの音が混在したものではなく，一つの音源からの音であれば，既に提案されている適応ピッチトラッキングの手法を用いることができるであろう[6]．しかし，そのような信号処理の手法は，そのままでは混合音に対して適用することはできない．そこで我々は，混合音に対して適用できる位相トラッキングの手法を考案した．これは，次の6ステップからなる．

(ステップ1) 入力信号 z に対して周波数解析を行い，基本周波数成分をすべて抽出する．z は複数の音源からの音の混合物かもしれないから，複数の基本周波数成分があるかもしれないことを考慮する．但し，複数の音源の基本周波数が整数倍の関係にあることは考えない．

(ステップ2) 抽出された各基本周波数について，対応する音源であるかもしれない波形 q_i を選び出す．ここで i は選び出された波形を数える添字である．以下，この波形 q_i を「位相トラッキング前の原テンプレート」と言う．

(ステップ3) 狭帯域の帯域フィルタを q_i に適用する．帯域フィルタの中心周波数 f_i は，それぞれの q_i の平均的な基本周波数とする．帯域フィルタの出力は，正弦波に近い波形となるので，その位相をバッファに保持する．位相の時系列を $p_{q,i}(k)$ とおく（k は時刻であり，添字の $_{q,i}$ は q_i に対する位相の時系列であることを表す）．

(ステップ4) q_i に対して適用したのと同じ帯域フィルタを入力信号 z に対して適用し，ステップ3と同様に位相 $p_{z,i}(k)$ を保持する．

(ステップ5) 入力波形とテンプレート波形の時々刻々の時間差 $\delta k_{q,i}(k)$ を求める．位相差 $\delta p_{q,i}(k)$ は

$$\delta p_{q,i}(k) = p_{z,i}(k) - p_{q,i}(k), \tag{5}$$

で与えられるから，時間差 $\delta k_{q,i}(k)$ は

$$\delta k_{q,i}(k) = \frac{f_s}{2\pi f_i}\, \delta p_{q,i}(k), \tag{6}$$

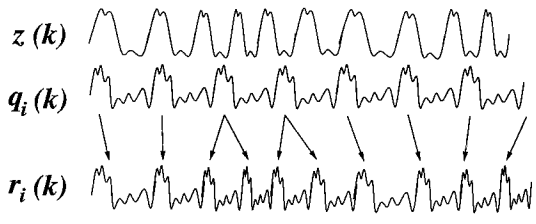
によって計算できる．ここで f_s はサンプリング周波数であり，$\delta k_{q,i}(k)$ の単位はサンプルである．

(ステップ6) 時刻 k における，位相トラッキングの処理結果となるべき波高値 $r_i(k)$ は，ステップ5で求められた時間差を用いて

$$r_i(k) = q_i(\, k - \delta k_{q,i}(k)\,) \tag{7}$$

によって求めることができる．以下，式(7)の r_i を「位相トラッキング後の原テンプレート」という．実際の処理では式(7)の r_i を式(4)の r_i として用いる．

図3は，上記のアルゴリズムが動作する様子を表した説明図である．



上段：入力波形 z；中段：位相トラッキング前の原テンプレート q_i；下段：位相トラッキング後の原テンプレート r_i．下段の波形 r_i がテンプレートフィルタリングに用いられる．なお本図は説明図であり処理結果を示したものではない．

図3 位相トラッキングの説明図
Fig. 3 Effect of phase tracking.

以上述べたように，本手法は，音源の変動を基本周波数の揺らぎと，基本周波数に対する高調波の相対位相や振幅の変動による波形のひずみとに分けて考え，前者を位相トラッキングによって，また後者をテンプレートフィルタリングによって吸収するものである．

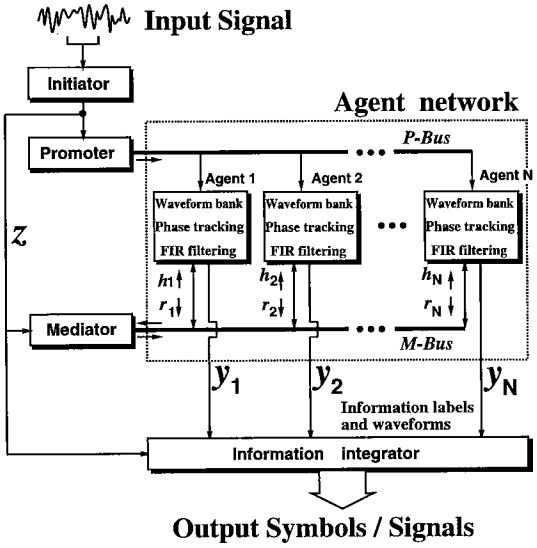
3. マルチエージェントアーキテクチャによる実装

本章では，前章で導入した計算を行うためのシステムの構成について議論する[7]．

3.1 概　　要

同時に複数の認識対象の音が存在し得るとき，ある音をシグナルととらえ他の音をノイズとみなすような処理モジュールを複数準備しておき，それらを並列に動作させることによって個々の音の認識を図るのは，極めて自然な発想であろう．それぞれの処理モジュールは，各々が担当する音だけを検出するという比較的単純な機能をもち，また処理モジュールは全く独立にではなく，相互に影響を及ぼしながら動作する．これはマルチエージェントアーキテクチャの考え方[8],[9]に他ならない．

図4に，提案するシステムの処理モデル（アーキテクチャ）を示す．このシステムは，複数種類の音が混在した音響信号を入力とする．本論文の範囲では，入力信号は音楽演奏である．出力としては，楽譜に類似した形式の記号表現および各音源ごとの音響信号を生成する．



本論文では，主に Agent network の部分を扱っている．

図4　提案する Ipanema アーキテクチャ
Fig. 4　The proposed "Ipanema" system architecture.

図4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）からなっている．そこで，図4のアーキテクチャを「Ipanema」と呼ぶ．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．

3.2 処理モジュール

3.2.1 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．

3.2.2 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．

3.2.3 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．これらの単音の波形は，「位相トラッキング前の原テンプレート」q_i（図3の中段の波形）として用いられる．

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「位相トラッキング後の原テンプレート」r_i を生成する．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 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 1 に示したテンプレートフィルタリングが実現される．

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．

M-Bus に関しては，現在の実装では，エージェントはメディエータに対してのみ情報を渡し，またメディエータからのみ情報を受け取る．しかし将来的には，M-Bus を介して任意のエージェントが任意のエージェントに対して情報を受け渡すような処理形態も考えられる．この場合 M-Bus は，黒板モデルにおける黒板の役割を果たしていると見ることもできる．

3.2.4 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，位相トラッキング後の原テンプレート r_i を読み込む．これらに基づいて，連立方程式 (4) を解けば，各エージェントに対するフィルタ係数 h_i が得られるので，これを M-Bus に書き込んでエージェントに返す．

3.2.5 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，情報インテグレータにおいて，音楽としての制約（文脈情報）を加味した上で音源の判定を行う．

具体的には，エージェントの最大相関のラベルだけを採択するのではなく，すべてのラベルを，それぞれの入力フレーム波形との相関値に基づいて評価された確信度をもつ仮説として扱う．次に，単音同士の時間的つながりを抽出して確率ネットワークをつくり，更にこのネットワークを利用して，単音同士のつながりを考慮した上での音源種類の確信度を計算する [10]．

なお，情報インテグレータの詳細については別稿にて報告する予定である．

4. 評価実験

提案法による音源同定の精度を評価するため，和音試料を用いたベンチマークテストと音楽演奏を用いた音源同定実験とを行った．

4.1 ベンチマークテスト

ここに述べるベンチマークテストは，文献 [4] で行ったものと同様のものである．用いたテストデータは，図 5 に示すような三つの単音からなる和音（3 和音）200 を並べた音響信号である．和音パターンはクラス 2 とした．クラス 2 とは，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（完全 5 度の音程にある）ようなパターンのうちで，基本周波数が倍音関係にある（同一，1 オクターブ，1 オクターブと完全 5 度，… の音程にある）ような単音の組を含まないようなパターン(注2)のことである [4]．このようなパターンでは，和音を構成する単音の音高を全くランダムに選定した場合に比べ，周波数成分の重複の程度が大きく，従って音源同定が難し

(注 2)：調律の仕方や演奏時の音程の揺れによって，楽譜上完全 5 度の音程を演奏した場合であっても，基本周波数の比は厳密には 1.5 になるとは限らない．例えば，ピアノを平均律で調律したとすれば，楽譜上完全 5 度の音程にある 2 音の基本周波数の比はおよそ 1.498 となる．しかし本実験では，調律や演奏の揺れによる基本周波数のずれは無視した．

くなる．これは，ランダムの場合に比べると，協和する単音から成り立っている多くの音楽に近い状況でもある．

パターンの作成においては，あらかじめフルート，ピアノ，およびバイオリンの自然楽器の単音を半音ごとにスタジオで収録した（16 bit, 48 kHz）．収録した波形を計算機上に蓄積し，クラス2およびMIDIノート番号(注3) 60～84という制約の中でランダムに選択して加算することによってパターンを作成した．なお，この音域は，フルート，バイオリン，ピアノのいずれもが発音可能であり，かつ通常よく用いられる音域として選定した．各和音は500 msの継続時間とした．

本実験では，あらかじめシステムに蓄積する波形（テンプレート生成に用いる波形）としてテストパターンの生成に利用するのと同一の波形を用いたり，同一個体の楽器を用いたりすると，波形の一致度が高いために評価実験としては適切でない．そこで，テンプレート生成に用いる波形とテストパターンの波形は，互いに異なる個体から収録したものを用いた．これを表1に示す．なお各楽器の奏者は，各楽器を専門とする音大生または音大卒業生であり，フルートは楽器個体ごとに1名で計2名，ピアノとバイオリンは，楽器個体に共通で各1名ずつとした．また，音源同定処理のみの精度を評価するため，各単音の音高と開始時刻は人手によりシステムに与えた（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）．更に，後処理モジュールである情報インテグレータの動作は停止させた（すなわち，単に入力フレーム波形

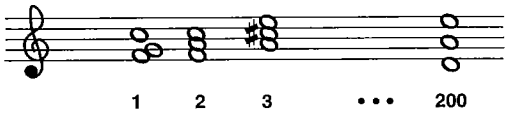


図5　ベンチマークテストに用いる和音パターンの例
Fig. 5　An example chord pattern used in the benchmark tests.

表1　ベンチマークテストに用いた楽器
Table 1　Musical instruments used in the benchmark tests.

	テンプレート	テストパターン
ピアノ	ペーゼンドルファ225	ヤマハC2
バイオリン	ハンニバルファグノラ	1720年クレモナ製
フルート	ブランネンクーパー	アルタス（頭部）＋ムラマツ

との最大相関値を与える波形を出力したエージェントのラベルを音源同定結果とみなした）．

本実験では音高については正解を与えているので，出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，音源同定精度 R は，単に

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

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ことを用いて，文献[4]で用いられた認識率の定義式を変形したものである．

予備実験の結果から，テンプレートフィルタリングOnの条件においては，FIRフィルタの次数を40とした．なおテンプレートフィルタリングOffとは，FIRフィルタの次数を1としたという意味である．

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．但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，全く発音しないこともあり得るとした）．従って，仮にシステムが全く音源同定能力をもたずランダムな結果を出したとすると，R の期待値は33.3[%]となる．

実験結果を表2に示す．表2では，右下の欄の条件（テンプレートフィルタリングOff，位相トラッキングOff）が，単純なマッチトフィルタによる音源同定に相当している．そこで，両者Offの場合と，表2の中の他の3条件との間で統計的検定を行った．その結果，位相トラッキングとテンプレートフィルタリングの両者がOnの場合には，有意水準0.1%で，音源同定精度の改善効果が有意となった．また，位相トラッキングがOnでテンプレートフィルタリングがOffの場合にも，改善効果は有意となった（有意水準1%）．

一方，位相トラッキングがOffでテンプレートフィ

表2　ベンチマークテストの結果
Table 2　Benchmark test results.

		テンプレートフィルタリング	
		On	Off
位相トラッキング	On	69.5% ± 3.8%	61.7% ± 3.6%
	Off	58.7% ± 3.6%	55.0% ± 3.4%

± は95%信頼区間を示す．

(注3)：MIDIノート番号とは音高を示す番号であり，中央ドを60とし，半音ごとに1ずつ異なった番号を与えたものである．

ルタリングがOnの場合には，有意水準10%としても改善効果は有意でなかった．そこで，位相トラッキングをOnに固定し，テンプレートフィルタリングがOnの場合とOffの場合とについて，改めて同様の検定を行ったところ，有意水準0.5%で，テンプレートフィルタリングOnの場合の方が音源同定精度が高いことがわかった．この結果は，2.2の冒頭に述べたように，テンプレートフィルタリングが位相トラッキングを併用することで初めて効果を発揮する手法であることを示していると言える．

4.2 音楽演奏を対象としたテスト

ベンチマークテストに加え，音楽の実演奏4曲を対象とした音源同定テストを行った．テスト曲の曲名を表3に示す．これらの曲はいずれもバイオリン，フルート，およびピアノの3パートからなり，各パートは単旋律となっている．例えば「蛍の光」は，文献[5]に掲載されている楽譜を演奏したものである．表3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，テンプレートフィルタリングOnの条件においては，予備実験の結果からFIRフィルタの次数を20とした．また，ベンチマークテストと同様，音源同定処理のみの認識率を評価するため，各単音の音高と開始時刻は人手によりシステムに与えた．各単音がバイオリン，フルート，ピアノの3楽器のうちのいずれかであることは既知とし，パート数や各楽器の同時発音数は未知とした．情報インテグレータでは，単音のつながりを考慮した処理は行わず，単に入力フレーム波形との最大相関値を与える波形を出力したエージェントのラベルを音源同定結果とみなした．

表4に実験結果を示す．表中の値は，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，それらを平均したものである．結果の定性的傾向はベンチマークテストと同様であり，提案手法の効果が示されている．なお，ベンチマークテストの結果との数値上の差が見られる原因としては，演奏に用いた楽器個体の違いや，音楽演奏では各単音の長さがまちまちであり，かつ音楽的表現が行われるために，一般に各単音の変動がベンチマークテストで用いたテストデータよりも大きいなどといった条件の違いが考えられる．

表3 音楽演奏を対象としたテストに用いた曲
Table 3 Music used in the experiments.

曲名	含まれる単音数
アニー・ローリー	234音
ローレライ	297音
旅愁	304音
蛍の光	242音

表4 音楽演奏を対象としたテストの結果（4曲の平均）
Table 4 Experimental results from tests on actual music.

		テンプレートフィルタリング	
		On	Off
位相トラッキング	On	67.8%	65.0%
	Off	63.2%	60.8%

5. むすび

本論文では，音楽のように複数の音源が混在した音響信号に対する音源同定を目的とし，実環境における音の多様性や変動に対応した方法として，適応型混合テンプレートを用いた音源同定処理を提案した．更に，提案手法の応用として，アンサンブル演奏に対する音源同定システムを構築した．自然楽器音の単音によるベンチマークテスト，およびアンサンブルの実演奏を用いた実験の結果，単純なマッチトフィルタによる音源同定処理に比べ，提案手法が音源同定精度を改善する効果をもつことが確認された．

従来，マルチエージェントアーキテクチャに基づく音響処理システム[2],[3]では，エージェント間における出力の調整は明確に定式化されないことが多かった．これに対し本論文のシステムでは，エージェントの出力の調整を2乗平均誤差の最小化という規範で定量化している点が特徴である．また，これまでに，複数種類の楽器のアンサンブル演奏を扱うことのできる音楽認識システムも提案されているが[4]，自然楽器による実演奏を精度良く認識することは難しかった．これに対し本論文は，アンサンブルの実演奏に対する認識の可能性を示したものと位置づけることができる．

しかし，本手法だけでは，まだ実用的な音源同定精度が得られているとは言えない．また，本手法では，処理精度がテンプレートフィルタリングの次数の設定やテンプレートバンクに蓄積する波形の組合せに影響を受けることがわかっている．従って，今後の課題として，これらを自動的に適切に設定する処理について検討する必要がある．また，FIRフィルタによるテンプレートフィルタリングでは，音の変動のすべてが吸

収できるわけではなく，例えば音のかすれのような非線形な変動に対しては吸収の効果が期待できない．このため，対象の多様性や変動をモデル化して，各音源のもつ変動の性質に基づいたフィルタを構成することは，高精度化を図る上で重要な課題である．

一方，我々は既に，後処理モジュールである情報インテグレータによって音源同定精度が改善されるという実験結果を得ている[10]．情報インテグレータの詳細については，稿を改めて報告する予定である．

謝辞 本研究に対しサポートして頂いたNTT基礎研究所の東倉洋一所長，石井健一郎情報科学研究部長，奥乃博主幹研究員，川端豪主幹研究員および中谷智広研究主任，音楽試料の収録に協力頂いた国立音楽大学の莱孝之助教授およびNTT基礎研究所の小坂直敏主幹研究員に感謝する．

文　献

[1] S. Amari and A. Cichocki, "A New Learning Algorithm for Blind Signal Separation," in Advances in Neural Information Processing Systems 8, MIT Press, pp.757–763, 1996.

[2] 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，" 知能誌，vol.12, no.1, pp.111–119, 1997.

[3] V. Lesser, S.H. Nawab, I. Gallastegi, and F. Klassner, "IPUS: An architecture for integrated signal processing and signal interpretation in complex environments," in Proc. of the 11th National Conf. on Artificial Intelligence, pp.249–255, 1993.

[4] 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-D-II, no.11, pp.1751–1761, Nov. 1996.

[5] 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-D-II, no.11, pp.1762–1770, Nov. 1996.

[6] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," IEEE Trans. on ASSP, vol.34, no.5, pp.1124–1138, 1986.

[7] K. Kashino and H. Murase, "A music stream segregation system based on adaptive multi-agents," in Proc. of the 15th Int'l. Joint Conf. on Artificial Intelligence, vol.2, pp.1126–1131, 1997.

[8] P. Maes, ed., "Designing Autonomous Agents," The MIT Press, 1990.

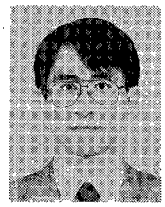
[9] 嘉数侑昇，"マルチエージェントシステムの研究動向，" システム/制御/情報，vol.41, no.8, pp.291–296, 1997.

[10] 柏野邦夫，村瀬　洋，"動的メロディー抽出を用いたアンサンブル演奏の音源同定，" 日本音響学会音楽音響研究会資料，vol.MA97-4, 1997.

(平成9年9月24日受付，10年1月12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．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および知識処理に興味をもつ．情報処理学会，日本音響学会，人工知能学会，IEEE各会員．



村瀬　洋　(正員)

昭53名大・工・電子卒．昭55同大大学院修士課程了．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．平4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．工博．昭60本会学術奨励賞，平4電気通信普及財団テレコムシステム技術賞，平6 IEEE-CVPR国際会議最優秀論文賞，平7情報処理学会山下記念研究賞，平8 IEEE-ICRA国際会議最優秀ビデオ賞受賞．情報処理学会，IEEE各会員．