音楽情報科学 19−11
(1997. 2. 21)

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 −

柏野 邦夫　　村瀬 洋

NTT基礎研究所
〒243-01 厚木市森の里若宮 3-1

kunio@ca-sun1.brL.ntt.co.jp, murase@apollo3.brL.ntt.co.jp

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．本稿では，音楽の生演奏など，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．さらに，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生演奏を用いた音楽認識テストを行った結果，単純なマッチトフィルタによる音源同定処理に比べ，提案手法が有効であることが確かめられた．

キーワード 聴覚的情景分析，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

Sound Source Identification Using Adaptive Template Mixtures
− Formulation and Application to Music Stream Segregation −

Kunio Kashino and Hiroshi Murase

NTT Basic Research Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi,
243-01, Kanagawa, Japan.

Abstract Sound source identification is an essential problem in auditory scene analysis when multiple acoustical objects are simultaneously present in the scene. However, little work has been done on sound source identificaiton for a multiple-source environment. Here we propose a novel method for sound source identification. The key idea is adaptation of templates, which has permitted to cope with variation of sounds. As an example application of the proposed method, we have built a music recognition system that recognizes instrument names and pitches of the notes included in ensemble music performances. Experimental results show that the proposed adaptive mechanisms significantly improve the accuracy of sound source identification in comparison to a conventional matched-filter-based method.

key words auditory scene analysis, sound source identification, sound source separation, music scene analysis, automatic music transcription, matched filter

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，その名の通り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，音によって周囲の状況を理解しようとする場合はもちろん，音声認識のように一見認識対象が決まっている場合であっても，実環境での自然なヒューマンインタフェースの手段として利用できるような完成度の高いシステムの構築を目指すのであれば，避けて通ることのできない問題である．

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．本稿では，前者を音源分離（sound source separation），後者を音源同定（sound source identification）と呼ぶ．従来の auditory scene analysis（聴覚的情景分析）の工学的研究では，主として音源分離だけが議論されてきた．一方，室内などの音響事象の認識 [1]，話者認識，音声区間の切り出しなど，単一で存在する音源の識別の研究も行われている．しかし，同時に複数種類の音が存在する状況下でそれぞれの音源を同定する問題は，これまでほとんど検討されていない．

そこで本稿では，音源同定の問題を扱う．音源同定に関する研究としては，柏野らによる OPTIMA の研究がある [2, 3]．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の音を用いて行われていた．これは，生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．

本稿では，多様で変動の大きい対象を扱うための鍵技術として「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簡単な評価実験を行って，**2.** で提案する処理の有効性を検討する．**5.** をむすびとする．

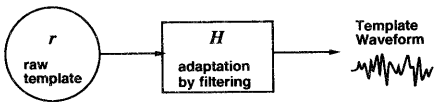


図 1: 音源波形のテンプレートフィルタリングモデル

2 適応型テンプレート

2.1 テンプレートフィルタリング

各音源波形の和で与えられる波形を各音源波形に分離する問題を解くための制約として，各音源波形のモデル $y_n(k)$ を与えることを考える．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，われわれの問題は，

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

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．音源波形 $y_n(k)$ 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，ある一群の音源波形を，原テンプレートと線形フィルタによる変形とでモデル化するものである．線形フィルタとして FIR 型を用いることにすれば，

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

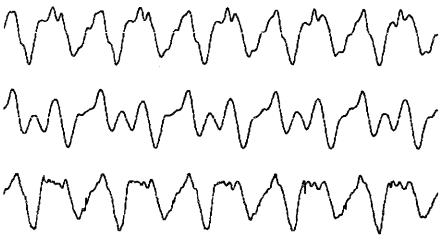
と書ける．ここで，h は FIR フィルタのインパルス応答，r は原テンプレート波形，M はフィルタの次数である．

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．したがって，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

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

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，全ての n と m に関して，偏微分 $\partial J/\partial h_n(m)$



上段はベーゼンドルファ，中段はヤマハのピアノである．どちらも同じ高さ (F4), 同じ時間部分 (立上りから 100ms ~130ms) 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．しかし，適切な FIR フィルタを通すことによって，中段の波形を上段の波形にあわせて変形させることができる．下段は，40 次の FIR フィルタによって変形させた中段の波形．上段の波形にかなり近づいている．サンプリング周波数は 48kHz.

図 2: ピアノ波形の多様性とその吸収

が 0 となることである．この条件を用いると，$N \times M$ 個の連立一次方程式

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

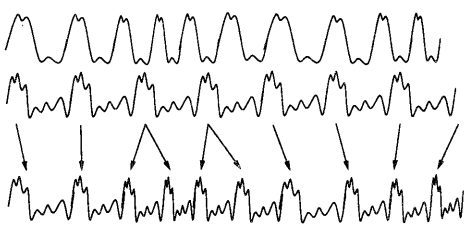
を導くことができる．未知数の個数と方程式の個数が等しいので，この連立方程式は，係数行列の逆行列を求めることによって解くことができる．式 (4) の係数行列は，$i \neq j$ のとき $r_i(k)$ が $r_j(k)$ の定数倍とならないように定めておけば，正則となる．

2.2 位相トラッキング

前節の処理が有効となるためには，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なくてはならない．なぜなら，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，r 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 [5]. 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．そこでわれわれは，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 ステップから成る．

(1) 入力信号 z に対して周波数解析を行い，基本周波数成分を全て抽出する．z は複数の音源からの音の混合物かも知れないから，複数の基本周波数成分があるかも知れないことを考慮する．ただし，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．



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；下段：位相トラッキング後の原テンプレート．下段の波形が式 (4) における原テンプレート $r_i(k)$ として用いられる．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

(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知れない原テンプレート r_i を選び出す．

(3) 狭帯域のバンドパスフィルタを r_i に適用する．バンドパスフィルタの中心周波数は，それぞれの r_i の平均的な基本周波数とする．バンドパスフィルタの出力は，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 $p_{r,i}(k)$ とおく（k は時刻）．

(4) r_i に対して適用したのと同じバンドパスフィルタを入力信号 z に対して適用し，**(3)** と同様に位相 $p_{z,i}(k)$ を保持する．

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

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

で与えられるから，時間差 $\delta k_{r,i}(k)$ は

$$\delta k_{r,i}(k) = \frac{f_s}{2\pi f_{c,i}}\, \delta p_{r,i}(k), \tag{6}$$

によって計算できる．ここで f_s はサンプリング周波数，$f_{c,i}$ は適用されたバンドパスフィルタの中心周波数である．

(6) 時刻 k での位相トラッキング後の波高値 $r_i(k)$ は，求められた時間差を用いて

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

によって求めることができる．

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．

以上述べたように，本手法は，音源の変動を基本周波数のゆらぎと，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歪みとに分けて考え，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．

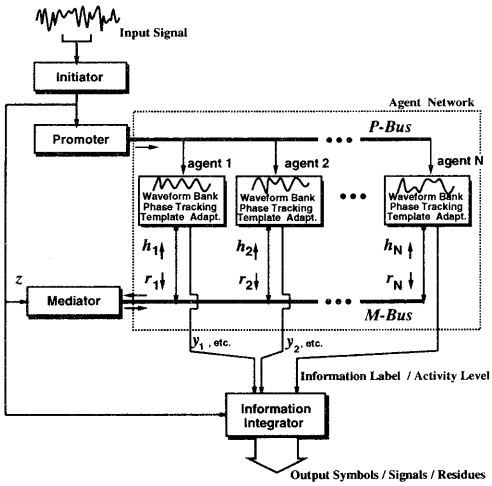


図 4: 提案する Ipanema アーキテクチャ

3 マルチエージェントアーキテクチャによる実装

本節では，前節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．

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捉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，きわ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．本稿の範囲では，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，各音源ごとの音響信号，および解釈の残差の音響信号を生成する．

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）から成っている．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，これをフレーム（ある時間範囲）ごとに切り出した波形を出力する．イニシエータの出力は，後続の処理のきっかけとなる．フレーム長は一定ではなく，音の立上りを検出するごとに新たなフレームを生成する．

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と，パワー包絡など音のおおまかな特徴量とを，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たち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決め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一つのエージェントは，個々の音源種類（例えばフルート）に対応している．エージェントは，原テンプレートの集合であるテンプレートバンクを持っている．テンプレートバンク中には，例えば半音ずつピッチの異なる単音の波形が原テンプレートとして蓄積されている．

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およびパワー包絡など単音のおおまかな特徴量が，自分の担当する音源種類の基本周波数および特徴量と矛盾しなければ，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，基本周波数や特徴量が現在の入力と最も整合する波形を選び出し，前節で述べた位相トラッキング処理を行って原テンプレート r_i を生成する．もし，P-Bus 情報中の基本周波数や特徴量が自分の担当する音源種類と矛盾すれば（たとえば発音不可能な音域であるなど），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原テンプレート 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る．メディエータからは，各エージェントに対応するフィルタ係数が返って来るので，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 **2.** に述べ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結果としての波形 y_i，活動レベルとしての平均パワー $E[y_i^2]$ 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．本稿においては，各エージェントの提案する原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，システムの出力の一部を修正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，活動レベル $E[y_i^2]$ および記号表現のラベルを受け取る．現在の実装は簡略化されていて，情報インテグレータはただ活動レベルに基づいて実際に発音している音源を判定するのみである．したがって現状では情報のインテグレーションという名前と実体とは整合していない．

しかし，近い将来このモジュールは拡張される予定である．これまでに述べた Ipanema の枠組では，フレーム単位に処理が行われるために散発的に発音判定の誤りが生じることがある．これに対しては，音どうしの継時的あるいは同時的関係に着目して出力判定を修正することがきわめて効果的である．この修正とは，結局，統計的情報など対象に関する複数の知識を統合して最も確からしい判断を行うことである．このようなわけで，本モジュールは情報インテグレータと名付けられている．情報のインテグレーションの具体的な手法としては，Bayesian network を応用した OPTIMA [2, 3] が一つのベースになり得ると考えている．

4 評価実験

提案法による認識精度を評価するため，単音認識のベンチマークテストと音楽認識テストを行った．

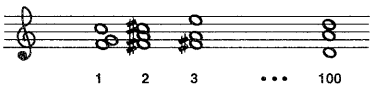


図 5: ベンチマークテストに用いる単音パターンの例

表 1: 実験に用いた楽器

	テンプレート	テストパターン
ピアノ	ベーゼンドルファ 225	ヤマハ C2
バイオリン	ハンニバル ファグノラ	作者不詳（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

4.1 ベンチマークテスト

ここで用いるベンチマークテストは，文献 [2] で行ったものと同様のものである．用いたテストパターンは，図 5 に示すような 3 つの単音が同時に鳴るパターンである．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一組が 1.5 の整数倍の関係にある基本周波数を持つような単音パターンのことである [2].

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（16bit, 48kHz）．この波形を計算機上に蓄積しておき，これをクラス 2 および MIDI ノート番号 60 ～74 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．

認識率 R の定義は，文献 [2] などと同様に

$$R = 100 \cdot \left(\frac{right - wrong}{total} \cdot \frac{1}{2} + \frac{1}{2} \right) \quad [\%], \qquad (8)$$

とした．ただし $right$ 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，$wrong$ 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，$total$ は入力 (正解) に含まれる総音符数である．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．

本実験では，原テンプレート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るたりすると，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，テンプレートの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．

表 2: ベンチマークテストの結果

		テンプレートフィルタリング	
		On	Off
位相トラッキング	On	77.3 % ± 4.1 %	64.7 % ± 4.9 %
	Off	61.0 % ± 4.5 %	57.8 % ± 4.8 %

± は 95 % 信頼区間を示す.

表 3: 音楽認識テストの結果

		テンプレートフィルタリング	
		On	Off
位相トラッキング	On	66.3 %	61.0 %
	Off	52.7 %	52.3 %

実験結果を表 2 に示す. この表では, 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. したがって, マッチトフィルタに比較して, **2.** で提案した適応型テンプレートを用いる処理の有効性が明確に示されていると見ることができる.

4.2 音楽認識テスト

ベンチマークテストに加え, 音楽の生演奏を対象とした音楽認識テストを行った. ここでは, 表 1 とはまた別の楽器個体のバイオリン, フルート, ピアノを用いて演奏したテスト曲「蛍の光」を対象として, 音源同定処理についての認識率 R を調べた. テスト曲の楽譜は [3] のものを用いた. 表 3 にその結果を示す. 表中の値は音源同定処理だけに関する認識率である. 結果の定性的傾向はベンチマークテストと同様であり, 提案手法の有効性が示されている.

5 むすび

本稿では, 音楽の生演奏など, 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した. さらに, 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した. このシステムは, マルチエージェントアーキテクチャに基づくことにより, モジュラリティとスケーラビリティを持たせた形で簡明に実装することができた. 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生演奏を用いた音楽認識テストの結果, 単純なマッチトフィルタによる音源同定処理に比べ, 提案手法が有効であることが確かめられた.

従来, マルチエージェントベースの音響処理システム [1, 4] では, エージェント間の通信は明確に定式化されないことが多かった. これに対し本稿のシステムでは, エージェント間の通信を二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 [2], 生演奏を扱うことは容易ではなかった. これに対し本稿は, 生のアンサンブル演奏に対する認識の可能性を示したものと位置付けることができる.

しかしながら, 認識精度自体はまだ実用的なものではない. 今後の課題として, まず情報インテグレータをフルに実装した形での評価が必要である. 次に, テンプレートフィルタのパラメトライズの問題がある. すなわち FIR フィルタだけではなく, 各音源のもつバリエーションをうまく表現するような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. このためには, 対象の多様性や変動をミクロにモデル化することも必要となろう. 一方でわれわれは, 適応型混合テンプレートの枠組を, 音楽以外の例題に対して適用することも試みたいと考えている.

謝辞

議論して頂いた, 弊社基礎研究所の奥乃 博 主幹研究員, 川端 豪 主幹研究員, 中谷 智広 研究主任に感謝する. また日頃サポートを頂く同研究所情報科学部の石井健一郎部長に感謝する.

参考文献

[1] Lesser V., Nawab S. H., Gallastegi I. and Klassner F. IPUS: An Architecture for Integrated Signal Processing and Signal Interpretation in Complex Environments. In *Proceedings of the 11th National Conference on Artificial Intelligence*, 249–255, 1993.

[2] 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, **J79-DII**, 11, 1751–1761, 1996.

[3] 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, **J79-DII**, 11, 1762–1770, 1996.

[4] 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, 12, 1, 111–119, 1997.

[5] Nehorai A. and Porat B.: Adaptive Comb Filtering for Harmonic Signal Enhancement. *IEEE Trans. on ASSP*, 34, 5, 1124–1138, 1986.