音楽情報科学 24－5
(1998. 2. 19)

アンサンブル実演奏の自動アンミキサ

柏野 邦夫　　村瀬 洋

NTT基礎研究所
〒243-0198 神奈川県厚木市森の里若宮 3-1

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

あらまし　従来、ステレオまたはモノラル収録の、複数楽器の混在した演奏をもとに、自動的に楽器ごとの信号に分離するような処理は、実現されていなかった。これは、複数の基本周波数成分の抽出や、それぞれの音がどの楽器で演奏されたものであるかの認識（音源同定）が困難なためである。これに対し本稿では、成分選択型分析合成法を提案し、これに複数の基本周波数を抽出する処理と、適応型混合テンプレートと単音連繋確率ネットワークを用いた音源同定処理とを組み合わせることによって、アンサンブル実演奏に対して自動的に音源分離を行うシステムが構成できることを述べる。なお、音源分離を行うシステムのことを、本稿ではアンミキサと呼ぶ。

キーワード　音源分離，音源同定，音楽情景分析，聴覚的情景分析，自動採譜

An Automatic Unmixer for Recordings of Real Ensemble Music Performances

Kunio Kashino and Hiroshi Murase

NTT Basic Research Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi,
243-0198, Kanagawa, Japan.

Abstract　Automatic unmixing for stereo or monaural music signals has been considered very difficult, because of the intractability of multiple-pitch extraction and sound source identification for real musical performances. Against this background, this paper proposes the selective analysis-synthesis method as a key technique to realize such an unmixer. This method is combined with a new multiple-pitch extraction method and a robust sound source identification method. A prototype system for an automatic unmixer for real ensemble music performances has been implemented and tested.

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

1 まえがき

複数の音源からの音響信号が混ざり合った音響信号をもとに、それぞれの音源の音響信号を復元する処理は、音源分離と呼ばれている。例えば、音楽演奏は、多くのパートの楽器や歌唱で組み立てられている場合が多いが、これを各パートごとの音に分けて取り出すような処理である。マルチメディア時代を迎え、我々の音楽への接し方も多様化してきているが、もし音楽に対する音源分離が実現できれば、音楽の楽しみ方が大きく広がるばかりでなく、音楽教育や音楽制作の場に革新をもたらすことも考えられる。

音源分離の研究は、四半世紀以上の歴史があるが、大きく分けて二つの考え方に分類することができる。その一つは、マイクロホン（入力チャンネル）を多数用意し、それら入力チャンネル間の情報を積極的に用いようとする考え方である。例えば、マイクロホンアレイによって集音の焦点を作るビームフォーミングの手法 [1] がこれにあたる。また、特に、音源の数と同数またはそれ以上の数のマイクロホンを用いることを条件とすると、問題は比較的簡単となり、古くから多数の手法が提案されているうえ、最近になってもなお、信号源の独立性に基づく方法 (Independent component analysis) が研究されている [2]。

音源分離の考え方のもう一つは、高々2チャンネルの情報で音源分離を実現しようとする考え方である。例えば、2つの耳で数多くの訴えを聞き分けたと伝えられる聖徳太子のような計算機を実現しようとする試みがこれにあたる [3]。我々が日頃 CD (Compact Disc) や放送等で接しているような音楽演奏を処理対象とした場合、利用できるのは高々2チャンネルであることから、この考え方は、音楽演奏を対象とする場合には、多数のマイクロホンの使用を前提とする方法よりも、より現実的な問題設定であると言えよう。

我々は、上記二つのうちの後者の立場から、通常のステレオまたはモノラル収録の音楽演奏の音源分離を検討している。複数楽器の混在した実演奏をもとに、楽器ごとの信号に分離するような処理は、従来は不可能であると考えられてきた。これは、複数の基本周波数成分の抽出や、それぞれの音がどの楽器で演奏されたものであるかを認識すること（この機能を音源同定と呼ぶ）が、非常に困難であるためである。

これに対し我々は、「適応型混合テンプレート」[4] と、「単音連繋確率ネットワーク」[5] を用いることによって、数種類の楽器音が混在したアンサンブル実演奏に対して、従来よりも高い精度で音源同定を行う方法を開発してきた。本稿では、この音源同定法と、新たに開発した基本周波数成分の抽出法および成分選択型分析合成法を組み合わせることによって、アンサンブル実演奏に対して自動的に音源分離を行うシステムが構成できることを述べる。なお、音源分離を行うシステムのことを、本稿ではアンミキサ (unmixer) と呼ぶ。

以下 **2.** では、システムの全体の処理の流れの説明と併せて、成分選択型分析合成法を提案する。**3.** において、アンサンブル実演奏に対する基本周波数抽出法について述べた後、**4.** において、適応型混合テンプレート法および単音連繋確率ネットワーク法による音源同定処理について説明する。**5.** で構築したシステムに対し動作を確認する実験を行って、**6.** をむすびとする。

2 成分選択型分析合成法

2.1 概 要

本稿で提案する手法の処理の流れを図1に示す。この処理は、基本的には正弦波加算合成 [6] に基づいている。すなわち、まず入力音響信号に対して高速フーリエ変換 (FFT) を行って短時間スペクトルを得た後、スペクトルのローカルピークを時間方向に接続することによって周波数成分を抽出する (Peak picking)。次に、抽出された周波数成分の中で、出力に用いるべき成分とその時間区間を選択 (Selection) し、選択された成分だけを用いて音響信号を再合成 (Sinusoidal synthesis) するという流れである。出力に用いるべき成分の選択においては、基本周波数抽出 (Pitch extraction) と音源同定 (Source identification) の結果を利用する。この方法を、成分選択型分析合成法とよぶ。従来から周波数成分を選択して合成する方法は提案されていたが、肝心の周波数成分の選択は人手に頼っていた。これに対し、成分選択型分析合成法は、成分の選択を自動的に行う方法である。

2.2 正弦波加算合成

McAulay と Quatieri は、音声（スピーチ）を多くの周波数成分の和で表現する分析合成法（以下 MQ 法という）を提案した [6]。この方法の概要は以下の通りである。まず、入力信号を、時間 T ずつ窓をずらしながら短時間フーリエ分析する。このとき、各フレームにおいて振幅スペクトルのローカルピークをピックアップする。すなわち、k 番目のフレームのフーリエ分析で求められる l 番目のローカルピークの振幅を A_l^k、角周波数を ω_l^k、位相を θ_l^k とする。次に、$k+1$ 番目のフレームにおける分析結果を参照して、両フレームでのローカルピークの対応づけを行う。文献 [6] では、

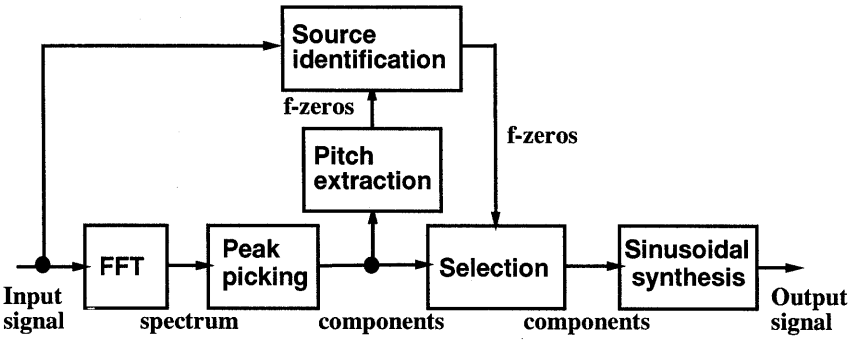


図 1: 成分選択型分析合成法に基づく自動アンミキサの処理の流れ

単に周波数の近接性に基づいてピークを対応づけているが、この対応づけには振幅値をも考慮するなど各種の工夫も可能である。ともあれ、いま、フレーム k における振幅 A_l^k、各周波数ω_l^k、位相 θ_l^k のローカルピークと、フレーム $k+1$ における振幅 $A_{l'}^{k+1}$、各周波数 $\omega_{l'}^{k+1}$、位相$\theta_{l'}^{k+1}$ のローカルピークとが対応づけられたとき、両フレーム間での位相の回転の様子を 3 次多項式で近似することを考える。すなわち、位相の近似値 $\theta_l(t)$ を、

$$\theta_l(t) = \theta_l^k + \omega_l^k t + \alpha t^2 + \beta t^3 \tag{1}$$

とする。ここで t は時刻を表し、フレーム k の中央を 0 とし、0 から T までの値をとる。このとき、α と β は次式で求められる（添字の l, l' は省略して記している）。

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \frac{3}{T^2} & -\frac{1}{T} \\ -\frac{2}{T^3} & \frac{1}{T^2} \end{pmatrix} \begin{pmatrix} \theta^{k+1} - \theta^k \\ -\omega^k T + 2\pi M \\ \omega^{k+1} - \omega^k \end{pmatrix} \tag{2}$$

ただし M は次の x^* に最も近い整数である。

$$x^* = \frac{1}{2\pi}\left\{\theta^k - \theta^{k+1} + (\omega^{k+1} + \omega^k)\frac{T}{2}\right\} \tag{3}$$

以上の計算を全てのローカルピークについて行えば、両フレームの間の波形は

$$s(t) = \sum_{l=1}^{L^k} A_l(t) \cos\theta_l(t) \tag{4}$$

として再合成できる。ここで L^k はフレーム k におけるローカルピークの数、$A_l(t)$ は振幅値を直線補完した値を表す。

2.3 混合音に対する正弦波加算合成

MQ 法はもともと、単一の音源を分析合成することを前提とした手法であり、混合音に対して用いる場合には、瞬時周波数の推定と、重複する成分の取り扱いに対して注意が必要である。

成分選択型分析合成において、本稿では、まず再合成すべき基本周波数成分を特定し、次に、その基本周波数成分と高調波成分とを選択する方法を用いる。基本周波数成分が特定されたとき、その高調波成分の選択は、それぞれの成分の瞬時周波数が整数倍かどうかを判定することで行う。したがって、瞬時周波数を正確に推定することは、高品質の分離結果を得る上で非常に重要である。

基本的には、各ローカルピークについて、式 (1) で求められる瞬時位相のフレーム間での平均変化率が、フレーム間における平均瞬時周波数に相当する。しかし、混合音を対象とする場合、各成分の位相が他の音源からの干渉を受けるので、計算される瞬時周波数が一時的に誤差の大きい値となることがある。この影響を避けるため、本稿では、一連のローカルピークに対して、フレームごとの瞬時周波数値の系列に対してメジアンフィルタリングを行って、誤差の大きい値を排除することとした。

さらに、音楽演奏では、異なる音源からの音であっても、高次の高調波成分は重複する場合がほとんどである。基本周波数成分が特定されれば、重複している成分も特定されるので、そのような成分については、再合成時に配慮する必要がある。例えば、図 2 に示すように、ある音 A と、A より完全 5 度高い別の音 B が混在しているとすると、A の第 3, 6, $\cdots$ 次高調波は B の成分と重複するため、それをこのまま A の音の分離結果として用いると、分離音の品質低下を招く可能性がある。

この対策として、(1) 重複成分をモデルから補完する、(2) 重複成分を直近の高調波成分から補完する、(3) 重複成分を、重複する音のうちのいずれか一つの合成のみに用いる（排他的割り当て）、などの方法が

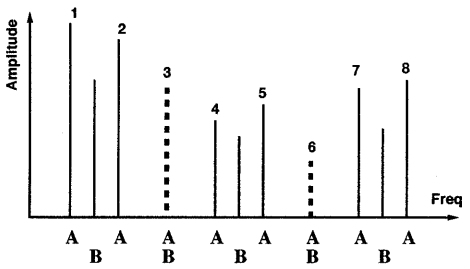
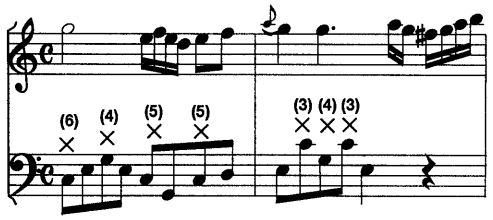


図 2: 重複する周波数成分の再合成時の扱いについて



× 印は、高調波の周波数が他パートの音の基本周波数と一致する音符を示す。括弧の中の数字は、基本周波数の比を示す。

図 3: アンサンブルにおける基本波の重複の例

考えられる。(1) は、A の音の具体的なモデルから重複成分の振幅と位相を取得する方法である。(2) は、A の第 n 次高調波の振幅と位相を、第 $n-1$ 次と第 $n+1$ 次の高調波の振幅と位相から合成する方法である。しかし、これらについて実験的検討を行ったところ、いずれも十分な分離品質を得ることは容易でないことが分かった。このため本稿では (3) の方法をとり、重複成分は、重複している音の中で最も基本周波数の高い音（図 2 の例では B の音）の合成のみに用いることとした。この方法は、数パート程度のアンサンブル演奏を対象とする限りでは、聴感上殆んど問題を生じないようである。

2.1 に述べたように、成分選択型分析合成法では、入力信号に含まれている基本周波数成分を特定する過程（基本周波数抽出）と、その基本周波数成分の音がどの楽器の音であるかを認識する過程（音源同定）が必要である。以下の章でこれらについて順次議論する。

3 基本周波数抽出

3.1 基本周波数抽出の問題点

従来、基本周波数抽出に関しては、音声の分野を中心に、零交差を用いる方法、ケプストラムを用いる方法、自己相関に基づく方法、線形予測分析に基づく方法、くし型フィルタに基づく方法など、非常に数多くの研究例があり、最近も研究が盛んである [9]。しかし、その殆んど全ては単一音を対象とする方法であって、複数音（複数の基本周波数成分）が含まれているかも知れない信号に対して、基本周波数成分の数と周波数とを抽出する方法は、サンプラの音のように変動の少ない音を対象とする方法 [8] を除いては、提案されていない。

アンサンブル演奏に対する周波数成分抽出の難しさとして、次の 3 点を挙げることができる。第 1 点は、周波数成分の重複である。アンサンブル演奏では、前章（図 2）で述べたように異なる音の高調波が重複するばかりではなく、あるパートの基本周波数が、他のパートの基本周波数のちょうど整数倍となっている場合が、意外に多い。一例として、Beethoven 作曲の 2 パートのアンサンブル曲 “Drei Duos” の第 1 曲冒頭の楽譜を図 3 に示す。図中で、低いパートの音符のうち × 印をつけたものは、同時に発音している高いパートの音符の基本周波数のちょうど整数分の一の基本周波数をもつ音符である。このような場合、上記の従来法をそのまま適用すれば、高いパートの音を抽出することができない。この問題点を換言すれば、同時発音数が定まっていない、ということでもある。従来研究では、有声区間では必ず基本周波数が一つだけある、ということを暗黙の前提としていたが、アンサンブル演奏を対象とする処理では、ある区間でいくつの基本周波数を抽出すべきかが自明ではない。仮に、パート数が事前に分かっていたとしても、休符もあり、また残響もあるので、あまり処理の助けにはならない。

難しさの第 2 点は、楽器音の特徴の変動が大きいという点である。基本周波数が仮に整数倍の関係にある複数音であっても、例えば高調波のパワー比などの特徴量が安定して抽出可能であるならば、各楽器音のスペクトルパターンのテンプレートを予め用意しておくことにより、それぞれの基本周波数を推定することも可能であろう。しかし、自然楽器の実演奏では、楽器の個体差、奏法、および収録条件等によってスペクトルパターンが非常に大きく変動するため、実際にはテンプレートを用いても精度良く基本周波数を推定することは難しい。

難しさの第 3 点は、音声に比べて音域が広いという点である。一般に基本周波数抽出においては、正しい周波数の整数倍あるいは整数分の一の周波数を抽出してしまうという誤り（倍ピッチ、半ピッチなど）が多いが、音声の場合には、人間が通常の発話で発声する音域を考慮すれば、後処理によってその種の誤りは

ほぼ修正可能である。しかし、音楽演奏では音域がはるかに広く、同様の修正は殆んど不可能である。

3.2 基本周波数抽出法

前項で述べた問題点から、本稿のシステムで、従来提案されている方法をそのまま用いることはできない。そこで、本稿では以下の方法をとった。

まず、入力音響信号を周波数解析し、ある時間区間ごとに周波数成分を抽出する。次に、各区間において、周波数成分をクラスタリングする。クラスタリングは、瞬時周波数の高調波関係からのずれと、振幅の同期的な変化からのずれを評価するような距離尺度を用いて、次のような手順で行う。

1. 最も低い平均周波数をもつ周波数成分をクラスタ中心 C_1 とする
2. 平均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 m_θ 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 m_θ を越えない周波数成分すべてを見い出し、それぞれのクラスタに所属させる

ここで m_θ は、別の音と判定するための距離のしきい値である。クラスタ中心となった周波数成分が、基本波成分と判定される。このクラスタリングは、基本的には文献 [8] の「単音形成処理」と同様の手順であるが、アンサンブル実演奏に対応できるよう、距離の定義が変更されている。すなわち、文献 [8] では、クラスタリングに用いる距離を、周波数成分の立上り時刻と平均周波数とに基づいて定義していた。しかし、これでは基本周波数が整数倍の関係にあるような単音を抽出できず、また残響などによる立上りの不明瞭化に対応できないないため、周波数成分の振幅の時間変化、および周波数成分の瞬時周波数の時間変化をより精密に参照することとした。

本方法の抽出精度は実験の章で測定するが、本方法は、基本的には距離の定義においてヒューリスティックスを含む方法であるから、今回用いた演奏に対してある精度を得たとしても、他の場合に同じ精度が得られるとは限らない。したがって、複数音が含まれているかも知れない音楽音響信号に対する正確な基本周波数抽出法は、今後も引き続き検討する必要がある。

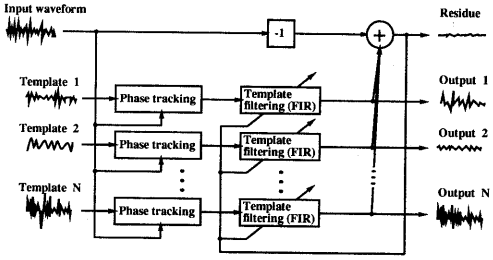


図 4: 適応型混合テンプレート法

4 音源同定

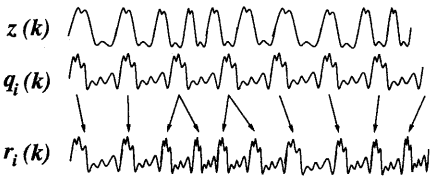
4.1 音源同定の問題点

音源同定の方法としては、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。しかし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 [7] 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 [8] 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一方、自然楽器の実演奏では、一般に楽器の個体差や演奏の表情づけなどによって楽器の音色が大きく変動するため、照合に基づく方法の精度にも限界がある。そこで本稿では、適応型混合テンプレート法と、単音連繋確率ネットワーク法とを組み合わせることによって、音源同定を行う。

4.2 適応型混合テンプレート法

適応型混合テンプレート法は、図 4 に示すように、各楽器音のテンプレートに対して、位相トラッキングと適応フィルタリングとを組み合わせて、入力混合音の波形を最も良く近似するようなテンプレート波形の組を求めることによって、入力に含まれる楽器の種類を同定する方法である。

適応型混合テンプレート法では、基本的には、マッチトフィルタのように、テンプレート波形と入力混合音の波形との相関値の大小によって、音源の有無を判別する。しかし、実際の楽器音は、常にテンプレート波形と相似な波形であるわけではなく、楽器の個体差や、演奏表現などによって波形は大きく変動する。そこで本手法では、音源波形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



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 5: 位相トラッキングの説明図

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。

まず、テンプレートの基本周波数および位相と、入力混合音の基本周波数および位相とを、時々刻々合わせ込む。この処理を位相トラッキングと呼ぶ。これは、テンプレート波形と入力音源波形とを、まず同じ中心周波数をもつ狭帯域バンドパスフィルタに通して基本周波数成分を抽出し、その位相を比較して、位相差に対応する時間だけテンプレート波形を遅延させることによって実現している。このアルゴリズムが動作する様子を図 5 に示す。

次に、高調波成分の振幅や位相のゆがみを吸収するための適応フィルタリングを行う。いま、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、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、

$$J = E\left[\left\{z(k) - \sum_{n=0}^{N-1} y_n(k)\right\}^2\right] \ , \qquad (5)$$

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 $y_n(k)$ は、位相トラッキングされたテンプレート (raw template) を FIR フィルタ H で変形させたものであるとすると、

$$y_n(k) = \sum_{m=0}^{M-1} h_n(m)\, r_n(k-m) \ , \qquad (6)$$

と書ける。ここで、h は FIR フィルタ H のインパルス応答、r は位相トラッキングされたテンプレート波形、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。ここでは、フィルタの係数 $h_n(m)$ を変えることを考え、J を最小とする $h_n(m)$ を求めると、$h_n(m)$ は $N \times M$ 個の連立一次方程式

$$\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} E\left[r_i(k-j)\, r_n(k-m)\right]\, h_n(m) = E\left[r_i(k-m)\, z(k)\right] \ . \qquad (7)$$

を解くことによって求めることができる (ここで $i = \{0,1,\cdots,N-1\}$, $j = \{0,1,\cdots,M-1\}$ である)。

4.3 単音連繋確率ネットワーク法

単音連繋確率ネットワーク法は、音のつながり（単音連繋）を考慮することによって、音源同定精度を向上させる手法である。本手法では、まず単音連繋を表現する確率ネットワークを作る。次に、前節の適応型混合テンプレート法によって求められた各テンプレートに対する相関値を、各音源の確からしさの初期値とした上で、統計情報を考慮して、確率ネットワーク上で各単音の音源確信度を更新し、最終的な音源同定結果を得る。

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 (8) によって定義される $Z(n_{k-1}, n_k)$ を考える。

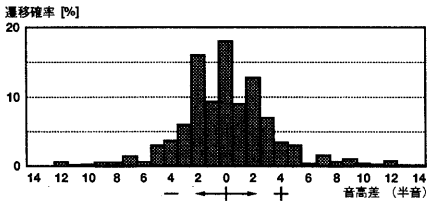
$$Z(n_{k-1}, n_k) = W \sum_i \Big\{ - \ w_i \log P_i(n_{k-1}, n_k) \Big\} \qquad (8)$$

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わ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つないだものを、単音連繋（music stream）と定義する。

なお W は時間窓に相当し、本稿では

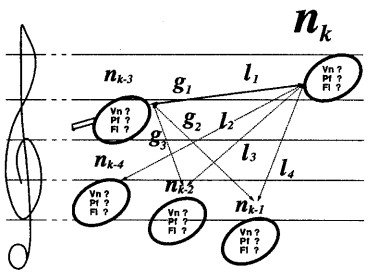
$$W(\delta t) = \exp\left(\frac{\delta t}{\tau}\right) \qquad (9)$$

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。



ポップスとジャズ計 397 曲の主旋律を分析した結果を示す。

図 6: 音高の遷移確率



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 7: 単音連繋ネットワークの作成の説明図

現在、単音連繋形成の要因としては、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 3 つを考慮している。音高遷移性とは、図 6 に示すように、旋律中に現れる音高の遷移確率を統計的分析によってモデル化したものである。また音色類似性とは、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たものである。さらに役割同一性とは、主旋律やベースラインなどといった音楽的な役割をもつパートが、時間的に一貫していることが多いことを評価したものである。

以上のように定義された Z 値を用いて、図 7 に示す手順によって単音連繋ネットワークを形成する。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している。このように形成された単音連繋ネットワークをベイジアンネットワークとして扱うことにより、音源確信度の更新を行うことができる [5]。

5 実験

5.1 基本周波数抽出の動作確認

まず、基本周波数抽出部の単体動作を確認するため、文献 [8] にあるものと同様のベンチマークテスト

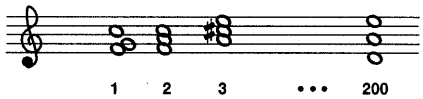


図 8: ベンチマークテストに用いる和音パターンの例

表 1: 基本周波数抽出実験の結果 [%]

	クラス 1	クラス 2	クラス 3
適合率	86.9	81.7	82.3
再現率	69.5	81.3	76.8

を行った。用いたテストデータは、図 8 に示すような 3 つの単音から成る和音（三和音）200 を並べた音響信号である。和音パターンはクラス 1、クラス 2 およびクラス 3 とした。クラス 1 とは、同時に発音する単音の少なくとも一組が整数倍の関係にある基本周波数をもつようなパターン（ただし同一の基本周波数である場合を除く）のことであり、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パターンのうちで、クラス 1 でないもののことである。またクラス 3 とは、クラス 1 でもクラス 2 でもないパターンのことである。

パターンの作成は、予め半音ごとに収録したフルート、ピアノ、およびバイオリンの自然楽器の単音を計算機上に蓄積し（16 bit, 48kHz）、クラス別および MIDI ノート番号 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。各和音は 500 ms の継続時間とした。システムにとっては、クラスや同時発音数は未知とした。

実験結果を表 1 に示す。表 1 で適合率とは、基本周波数抽出部が出力した基本周波数のうちの正解の割合を示し、再現率とは、入力和音に含まれていた単音のうち基本周波数抽出部から出力されたものの割合を示す。なお単音の正誤は MIDI ノート番号で判定した。距離のしきい値 m_θ の設定によって適合率・再現率の値は変化するが、表 1 は m_θ をある単一の値に固定して得たものである。

5.2 音源同定の動作確認

次に、音源同定部の単体動作実験を行った。音源同定部は、単音連繋を利用する部分を含むので、前項のようなベンチマークテストではなく、実際の音楽演奏を用いて実験を行った。用いた曲を表 2 に示す。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。標本化周波数は 48 kHz、量子化精度は 16 bit、2 チャンネルステレオ

表 2: 実験に用いたテスト曲

曲名	使用楽器（上のパートから順）	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

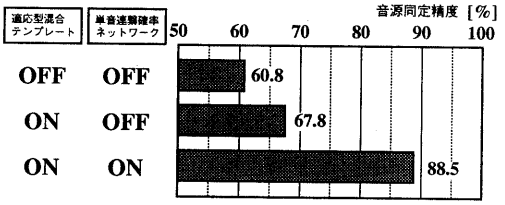


図 9: 音源同定実験の結果

入力とした。ただし音源位置の情報は全く使用していない。本実験では、各単音の基本周波数は人手で正しい値を与えた。

図 9 に実験結果を示す。これは、表 2 の各曲に対する音源同定精度を求めた後、それらを平均した値を示す。

5.3 システム全体の動作確認

前項と同じ入力音響信号（表 2）を用いて、音源分離実験を行った。本実験ではシステムを全自動で動作させた。

本実験の結果、聴感上は、各パートの周波数成分のうちの消え残ったものが聞こえ、また異音も聞こえるが、マイナスワン等の用途にはおおむね使用できる程度の出力が得られることが分かった。

なお、定量的評価の方法としては、例えば、アンサンブル演奏において各パートを別々に収録しておき、これを計算機上でミキシングしたものをテスト入力として、分離出力とミキシング前の各パートの信号とのスペクトル歪みを測定する方法が考えられる。しかし、アンサンブル演奏では視線や体の動きなどを使って各パートの奏者がタイミングを合わせることが多く、完全に各パートを別々に収録すると不自然な演奏になりやすい。また、聴感上の「分離の良さ」とスペクトル歪みとは、必ずしも対応しない場合がある。このようなことを考慮しつつ、本システム全体の定量的評価の方法については、今後引続き検討したい。

6 むすび

本稿では、成分選択型分析合成法に基づいて、モノラルまたはステレオで収録されたアンサンブル実演奏の音響信号に対して自動的に各パートごとの音響信号を分離して出力する「アンミキサ」を構築し、動作を確認する実験を行った。検討した技術的課題は、周波数成分の重複が含まれる場合の正弦波加算合成法、複数含まれているかも知れない基本周波数成分の抽出法、および音色の変動の大きい実演奏に対する音源同定法、の 3 点である。構築したシステムの動作実験では、おおむね良好な分離結果を得ることができた。

今後は、基本周波数抽出の高精度化について検討するとともに、高調波選択の高精度化や異音の抑制について検討を進める。また、システム全体の動作について、改めて評価実験を行って報告する予定である。

謝辞

ご指導頂くＮＴＴ基礎研究所の 東倉 洋一 所長、同研究所情報科学研究部の 石井 健一郎 部長、同研究部の 奥乃 博 主幹研究員，川端 豪 主幹研究員、および 中谷 智広 研究主任に感謝する。

参考文献

[1] 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).

[2] Lee T. and Orglmeister R.: "A contextual blind separation of delayed and convolved sources", *Proc. ICASSP 97*, 2, 1199–1202 (1997).

[3] 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).

[4] 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, 21–26 (1997).

[5] 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", 音響学研資, MA97-4, 23–28 (1997).

[6] McAulay R. J. and Quatieri T. F.: "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Trans. ASSP*, 34, 4, 744-754 (1986).

[7] 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).

[8] 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).

[9] 河原 英紀, de Cheveigne A.: "原理的に抽出誤りの存在しないピッチ抽出法とその評価について", 信学技報, SP96-96, 9–18 (1997).