楽器音の音響的特徴を知識として利用した選択的分離抽出法

窪　正晃　　鵜木　祐史　赤木　正人

北陸先端科学技術大学院大学　情報科学研究科

〒923-1292 石川県能美郡辰口町旭台 1-1

E-mail:　{kubomasa,unoki,akagi}@jaist.ac.jp

あらまし　本論文では、楽器音の音響的特徴を知識として利用し、複数の楽器音が混在する音の中から目的音を選択的に分離抽出する方法を提案する。提案法は、人が聴きたい音に“聞耳をたてる”とき、既知の目的音情報を利用しているという仮定に基づいたものである。提案するモデルは、目的音の音響的特徴を反映したテンプレートを利用して目的音の選択を行うトップダウン処理と、混合音成分から目的音の成分のみを分離するボトムアップ処理で構成されている。ここでは、（1）目的音に雑音が付加された状況、（2）二楽器音が混合した状況、（3）四種類の楽器音が混合した状況から、それぞれの目的音を選択的に分離抽出するシミュレーションを行った。この結果、提案するモデルが混合音中から目的音を選択的に分離抽出できることが示された。

キーワード　カクテルパーティ効果、計算論的聴覚情景解析、選択的分離抽出、二波形分離モデル

A method of selective segregation using knowledge sources as acoustic features of musical instrument sounds

Masaaki KUBO, Masashi UNOKI, and Masato AKAGI

School of Information Science, Japan Advanced Institute of Science and Technology

1-1 Asahidai, Tatsunokuchi, Nomigun, Ishikawa, 923-1292 Japan

E-mail:　{kubomasa,unoki,akagi}@jaist.ac.jp

Abstract　This paper proposes a method for selectively segregating the target sound from mixed musical instrument sounds using its acoustic features. This method is based on an assumption that one can use well-known information of the target sound to selectively segregate it from the mixed sounds. The proposed model is composed of the top-down processing for selecting the target sound using the template based on its acoustic features and the bottom-up processing for separating components of the target from components of the mixed sounds. Simulations were performed to selectively segregate the target sound from the mixed sounds, in the three situations: (1) target in the background noise; (2) target added to the other musical instrument sound; and (3) target added to the mixed three sounds. The results show that the proposed model can selectively segregate the target from the mixed sounds.

Keyword　Cocktail party effects, Computational Auditory Scene Analysis, Selective segregation, Auditory segregation model

1. はじめに

人間は混在する多くの音の中から、聴きたい音を選択し聴き分ける事ができる。この能力はカクテルパーティ効果[1]と呼ばれ、人間の聴覚システムが持つ優れた能力の一つとして知られている。人間は外界に生じる様々な音をカクテルパーティ効果などの聴覚能力により区別して聴き、それぞれの事象を把握し知覚する情報処理機能をもつ。このような聴覚による情報処理機能を計算論的アプローチから解明し、その優れた能力を計算モデルとして実現し工学的に利用することを目的とした研究が行われている。これらの研究は、計算論的な聴覚の情景解析(CASA: Computational Auditory Scene Analysis)と呼ばれている。

近年 CASA では、混在する音の中から目的音を分離抽出するという音源分離問題を計算機上に実装するこ

とを目的とした多くの研究が行われているが、複数の音源から目的音を選択的に分離抽出するという問題を完全に実現するには至っていない。その問題点としては、（１）混合音中の複数の音源はすべて分離抽出の対象となる可能性があるため、それらの中から目的音を選択し分離抽出を行う必要があること、（２）目的音の成分と他の音源の成分が周波数領域で重複する場合、重複周波数成分から目的音の周波数成分のみを分離する必要があること、があげられる。これらの問題に対し、従来の音源分離モデルは大きく分けてトップダウン処理に基づくものとボトムアップ処理に基づく手法が提案されている。

トップダウン処理に基づくモデルには、心理音響学的なグルーピング規則などを利用するモデル[2]や、マルチエージェントシステムを用いたモデル[3]があり、それらは複数の音源から目的音の選択が可能である。しかし、音響的な特徴として振幅（もしくはパワー）スペクトルを利用しているため、これらのモデルでは、二つの信号成分が同一周波数領域にある場合、完全に分離できているとは言い難い。一方、ボトムアップ処理として、鵜木,赤木の二波形分離モデル[4]が報告されている。これは、上記の問題（２）に対し、二波形分離問題を定式化し、Bregman が提唱した四つの発見的規則[5]を制約条件として利用することで、目的音と雑音の周波数成分の分離を可能にしている。しかし、二つの音源がどちらでも認識対象となりえるような問題設定ではないことから、上記の問題（１）には対応しきれていない。

そこで本論文では、複数の音源が混在する音響信号として音楽音響信号を対象とし、目的楽器音の選択的な分離抽出を行う方法を提案する。また、選択的な分離抽出を実現するために、音源間の重複周波数成分の分離が可能なボトムアップ処理（二波形分離モデル）と、目的楽器音のパワーやスペクトル形状などの音響的情報を知識として利用し抽出を行うトップダウン処理を融合した音源分離モデルを提案する。

提案法をベースに、様々な音に対する選択的分離抽出のモデル化を実現できれば、音声を対象にした場合には話者を選択した認識システムのフロントエンドへの応用、楽器音を対象にした場合には自動採譜や計算機音楽などへと、幅広く応用が期待できる。

2. 選択的分離抽出法

私達が日常生活の中で混合音の中から目的音を選択的に聴こうとするとき、その音をよく知っている場合や、直前に一度聞いておいた場合、比較的容易に目的音を聴き取ることができることがある。

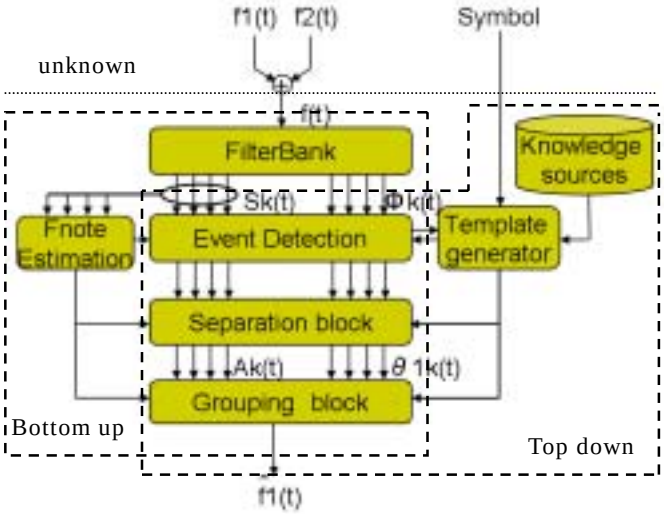

図１　選択的分離抽出モデルの構成

1. 周波数分析部：混合音の周波数分析（瞬時振幅と瞬時位相の抽出）を行う。
2. 特徴量テンプレート生成部：Symbol 入力に従い知識源より目的音に対応したテンプレートを生成する。
3. 基本周波数推定部：基本周波数を推定する。
4. 目的音選択部：混合音の周波数成分、基本周波数、特徴量テンプレートを用い目的音を選択する。
5. 波形分離部：混合信号の周波数成分から目的音のものと他音源のものを分離する。
6. グルーピング部：分離された目的音の周波数成分をまとめ、それらを再合成し目的音の音響信号を出力する。

図２　選択的分離抽出の処理の流れ

本論文では、この経験的な事実から、“人が聴きたい音を聴き取るとき、目的音に関する何らかの情報を利用しているのではないか”と考え、聴きたい音を聴く“聞耳”を表現するために次のような二つの仮定を行った。

1. 混合音中のどこかに目的音が存在することをあらかじめ知っている。

2. 目的音の音響的特徴を知識として持っている。

この仮定に基づき目的音の音響的特徴を知識として利用した選択的分離抽出法を提案する。実装した選択的分離抽出モデルの構成を図１に、モデルの処理の流れ図を図２に示す。モデルの入力は、観測された混合信号と仮定1で与えられる目的楽器名の Symbol である。また、仮定２により楽器音の音響的特徴を知識源として与えている。

この分離モデルは図中の破線で示すように、混合音中の目的音の位置等を選択するトップダウン処理と、目的音とその他の音源の周波数成分を分離するボトムアップ処理を融合したものである。ボトムアップ処理部には目的音とその他の音源が周波数領域で重複する場合でも分離が可能な二波形分離モデル[4]を利用している。しかし、このモデルは入力を二波形に限定しているため、本論文で取り扱うような複数の音源波形が混合した信号には直接利用できない。そこで、トップダウン処理部において入力された楽器音名と知識源から目的音に対応した特徴量テンプレートを生成し、さらにそのテンプレートを用いて混合音中から望みの楽器音成分を見つけ出す。これにより、混合音は選択された目的音波形と、それ以外すべての音源波形の二波形とみなすことができ、ここでも以下に示す二波形分離問題の定式化[4]に従って解くことができる。

はじめに、目的音の $f_1(t)$と、それ以外のすべての音源の混合音 $f_2(t)$が加算され、混合音 $f(t)=f_1(t)+f_2(t)$ のみを受音できるものとする。これを K 個の分析フィルタ群により周波数分解したとき、k 番目の分析フィルタを通過した $f_1(t)$と $f_2(t)$ の周波数成分をそれぞれ

$$X_{1,k}(t) = A_k(t)\exp(j\omega_k t + j\theta_{1k}(t)) \quad (1)$$
$$X_{2,k}(t) = B_k(t)\exp(j\omega_k t + j\theta_{2k}(t)) \quad (2)$$

と仮定すれば、$f(t)$の通過成分 $X_k(t)$ は

$$X_k(t) = X_{1,k}(t) + X_{2,k}(t) = S_k(t)\exp(j\omega_k t + j\phi_k(t)) \quad (4)$$

と表される。ただし、ω_kは分析フィルタの中心角周波数、$A_k(t)$, $B_k(t)$, $S_k(t)$は瞬時振幅、$\Phi_k(t)$は瞬時出力位相、$\theta_{1k}(t)$と $\theta_{2k}(t)$は瞬時入力位相である。ここで瞬時振幅 $S_k(t)$と出力位相 $\Phi_k(t)$は、それぞれ

$$S_k(t) = \sqrt{A_k^2(t) + 2A_k(t)B_k(t)\cos\theta_k(t) + B_k^2(t)} \quad (5)$$

$$\phi_k(t) = \arctan\left(\frac{A_k(t)\sin\theta_{1k}(t) + B_k(t)\sin\theta_{2k}(t)}{A_k(t)\cos\theta_{1k}(t) + B_k(t)\cos\theta_{2k}(t)}\right) \quad (6)$$

で求められるため、瞬時振幅 $A_k(t)$と $B_k(t)$ は、

$$A_k(t) = \frac{S_k(t)\sin(\theta_{1k}(t) - \phi_k(t))}{\sin\theta_k(t)} \quad (7)$$
$$B_k(t) = \frac{S_k(t)\sin(\phi_k(t) - \theta_{2k}(t))}{\sin\theta_k(t)} \quad (8)$$

として解くことができる。ただし $\theta_k(t)=\theta_{1k}(t)-\theta_{2k}(t)$ であり、$\theta_k(t)\neq n\pi, n\in\mathbf{Z}$ とする。

最後にすべての分析フィルタについて、瞬時振幅 $A_k(t)$と瞬時位相 $\theta_{1k}(t)$を求め、式(1)により周波数成分を合成することで目的音 $f_1(t)$を再合成することができる。

3. 選択的分離抽出モデル

提案するモデルは（１）周波数分析部、（２）基本周波数推定部、（３）特徴量テンプレート生成部 、（４）目的音選択部、（５）波形分離部の五つの処理ブロックで構成される。各処理ブロックの実装など詳細な点を以下で述べる。

3.1. 周波数分析部

分離の対象である楽器音は、弦や管などの振動や共鳴により発生する音であり、調波性が重要な特徴である。そこで、これらの情報を有効に利用するために、周波数分析処理部を狭帯域の定帯域幅フィルタバンクとした。また、楽器音はかなり高い周波数成分を含むため、サンプリング周波数を 20 [kHz]として、解析周波数範囲を 0～10 [kHz]、フィルタの帯域幅を 20 [Hz]（500 チャネル）として周波数成分の重なりを抑えた。

3.2. 基本周波数推定部

基本周波数推定部にはフィルタバンクの出力を利用した周波数軸上での Comb filtering [7]による手法を用いた。この方法は比較的雑音に対してロバストであり、また、信号の調波性を利用する手法であるので楽器音の情報を有効に利用できると考え採用した。

3.3. 特徴量テンプレート生成部

特徴量テンプレート生成部では、目的音の選択と目的音波形の分離を行う際に利用するテンプレートして、目的音の調波成分の振幅包絡形状と、各調波成分のパワー比を反映した時間周波数平面上の形状を生成する。特徴量テンプレートの生成には、目的音候補の基本周波数と持続時間、さらに知識源として持っている各楽器の特徴を利用する。ここで、知識源に蓄えられている各楽器の特徴量は、楽器音の同定手法[7]にも利用されているもので、本論文で楽器音を周波数分析した瞬時振幅を用いている。各楽器音の特徴を以下に示す。

- フルートなどの木管楽器の特徴
  - i. 各調波の振幅包絡の定常部分が平坦である。
  - ii. 高調波成分の数が 7 次以下である（周波数方向での減衰が大きい）。
- ヴァイオリンなどの弦楽器の特徴
  - i. 各調波の振幅包絡が激しく変動する。
  - ii. 高調波成分の数が多い（20 次程度）。
  - iii. 高調波成分では FM 成分が見られる。
- ピアノなどの打弦楽器の特徴

i. 各調波の振幅包絡に定常部分がなく、立上り後すぐに減衰する。
ii. 高調波成分が比較的多い（10 次程度）。
iii. 各調波成分の立ち上がりが鋭い。

各楽器のテンプレートは、上記の特徴をもとに、まず基本波にあたる調波成分の振幅包絡形状から作成される。このとき、振幅包絡は比較対象となる目的音候補の持続時間に応じて変化させている。その後で、既に得られた基本波の振幅包絡を、楽器によって異なる周波数軸上でのパワー比に従って修正し、各調波成分へ並べることでテンプレートを完成させる。以上の処理に基づいて作成されたフルート、ピアノ、ヴァイオリンに対する標準的なテンプレートの形状を図 3 に示す。

3.4. 目的音選択部

目的音選択処理部ではこれまでに得た情報から混合音中に含まれるすべての楽器音を目的音候補として抽出し、最終的には各候補と特徴量テンプレートの比較を行うことで候補から目的楽器音成分を選択する。この処理の流れを以下に示す。

1. 推定された基本周波数の周波数軸に対するヒストグラムから、出現頻度の高い周波数を混合音中の基本周波数候補として抽出する。
2. それぞれの基本周波数候補に対応する分析フィルタの瞬時振幅から時間領域での立上り、立下りを求める。
3. 混合音の周波数成分から 1.で求めた各基本周波数候補の整数倍の成分のみを 2.で求めた持続時間で抽出し目的音候補とする。
4. 3.で得られた目的音候補と特徴量テンプレートとの相関を求め、最も相関の高いものを目的音として出力する。

以上のような基本周波数の整数倍の高調波成分を抽出する単純な処理で得られた目的音の周波数成分には、従来のトップダウン処理モデルで問題となっていた他音源の重複周波数成分が存在している。そこで、波形分離部において、重複周波数成分の分離を行い、目的音を精度よく抽出する。

3.5. 波形分離部

波形分離部では、選択された目的音に残っている他音源の周波数成分を分離する処理を行う。具体的には、二波形分離モデルで利用している Bregman の発見的規則(iv)「一つの音響事象に属する共通の変化に関する

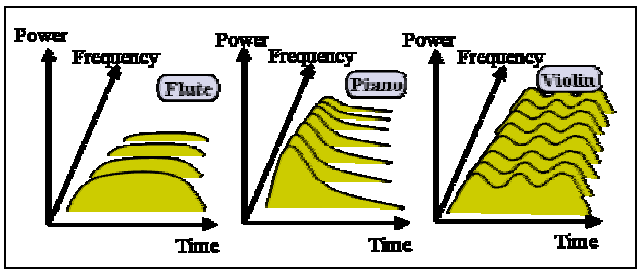

図 3　各楽器の標準的なテンプレート

規則」に対応する制約条件に注目する。二波形分離モデルでは、発見的規則(iv)を式(9)で表される制約条件としてとらえ、各調波成分における瞬時振幅 $A_k(t)$と隣接する調波成分の振幅包絡 $A_l(t)$の相関が最大となるように、各調波成分における $S_k(t)$と $\Phi_k(t)$から $A_k(t)$と $\theta_{1k}(t)$を求めている[4]。

$$\frac{A_k(t)}{\|A_k(t)\|} \approx \frac{A_l(t)}{\|A_l(t)\|}, \qquad k \neq l \qquad (9)$$

提案法でも同様にこの制約条件を利用するが、式(9)の $A_l(t)$を特徴量テンプレートの振幅包絡として、分離抽出する目的音の瞬時振幅とテンプレートとの相関が最大となるときの $A_k(t)$と $\theta_{1k}(t)$ を求める。このとき、テンプレートとの相関は、式(9)のような大きさを正規化した相関ではなく、大きさも考慮した相関にすることで各調波成分のパワー比による特徴も、分離抽出の処理に反映する形とした。

4. シミュレーション

選択的分離抽出モデルを評価するため、対象とする楽器音を、通常の演奏法による単音として、次のような条件で分離抽出シミュレーションを行った。

(1) 目的音に白色雑音を付加した信号
(2) 二つの楽器音を混合した信号
(3) 四種類の楽器を混合した信号

まず、(1),(2)のシミュレーションにより、提案法がこれまでよく考えられてきた音源分離問題に対しても有効な処理であることを確認する。特に(1)では、周波数領域の全帯域に雑音が存在する状況での分離精度を検証し、(2)では二楽器音の高調波における重複成分の分離効果を検証することで、提案法の有効性を確認する。次に(3)のシミュレーションにより、単音の楽器音ではあるが、実環境に近い状況で、提案法が選択的分離抽出を可能としているかどうかを検証することで、有効性を示す。

シミュレーションで利用する混合信号は、目的音と雑音の SNR を-10 dB から 20 dB まで 10 dB 刻みで変化させた信号とした。モデルの評価は、次式に示すよう

な SNR を利用して行った。この評価尺度は、元信号 $f_1(t)$ と選択的に分離抽出された信号 $\hat{f}_1(t)$ との差を雑音とみなした場合の時間領域での SNR である。

$$\mathrm{SNR} = 10\log_{10}\frac{\int_0^T f_1(t)^2\,dt}{\int_0^T (f_1(t)-\hat{f}_1(t))^2\,dt}\ [\mathrm{dB}] \qquad (10)$$

また、提案法の有効性を明確にするために、提案法以外の二つの処理方法との結果と比較した。

(a) トップダウン処理のみの結果：目的音の基本周波数の整数倍の成分を単純に抽出し、波形分離処理なしで再合成したもの。

(b) ボトムアップ処理のみの結果：波形分離部において特徴量テンプレートを利用せずに行った結果。

4.1. 雑音中からの分離抽出

まず、白色雑音が付加された状況からの目的音の分離抽出精度を評価した。目的音のすべての調波成分に雑音が存在する状況であるので、波形分離効果を検証できる。また、雑音下からの分離を行う従来の音源分離モデルとの比較も可能である。音楽音響信号を考えたときには、シンバルなど破裂音を生ずる打楽器を含む信号を模擬していると考えられる。

目的音はフルートとし、四種類の音高について分離抽出した結果の平均を図 4 に示す。横軸が入力の SNR であり縦軸が推定精度を表す SNR である。この結果、トップダウン処理のみの結果と比較することで提案法の波形分離処理の効果が確認できる。本論文では、狭帯域のフィルタバンク(20 [Hz], 500 チャネル)を使用しているため、各調波成分に残る雑音のパワーが少なくなる。そのため、入力の SNR が高い場合にはあまり差がないが、入力の SNR が低い場合に分離効果が明確に見られる結果となった。また、ボトムアップ処理のみの結果と比較すると、すべての雑音レベルで提案法が良い結果となっている。このことから、分離処理にテンプレートを利用することの効果が確認できる。

4.2. 二楽器混合音からの分離抽出

次に二種類の楽器音が混合された状況での分離抽出シミュレーションを行った。この混合信号では目的音候補が二つ存在する状況で、提案法の目的音選択部によって目的音を正しく選択可能であるか検証した。混合した楽器音はヴァイオリン(C4)とフルート(A4)とし、目的音をフルートとした。結果を図 5 に示す。この結果、二つの認識対象を持つ混合音に対して目的音選択処理部での処理により目的音を正しく選択できた。

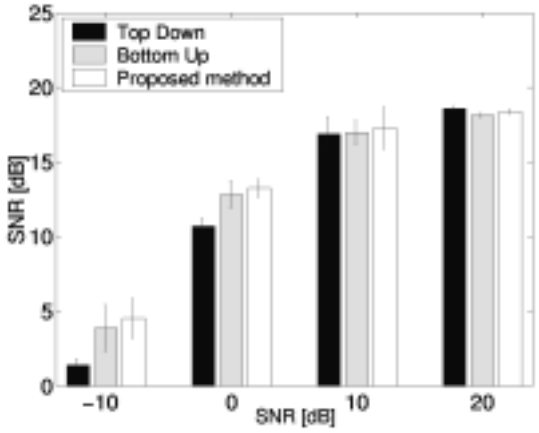

図 4　雑音中からの分離精度比較

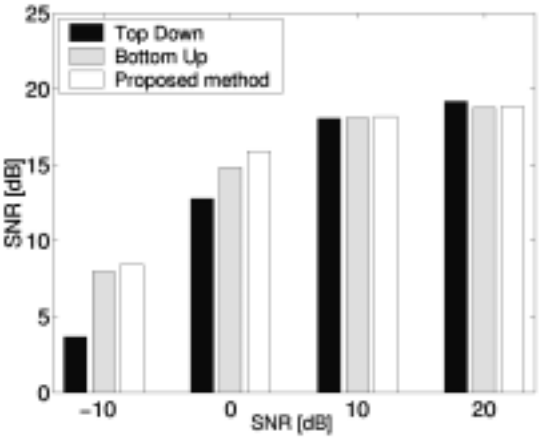

図 5　二楽器混合音からの分離精度比較

また、この音階の組み合わせでは目的音の高調波成分の 3,6,…次が重複する成分であるが、波形分離部での処理による分離効果から SNR が改善されている。特に、入力の SNR が低い場合には大きな効果が見られる。

4.3. 複数楽器混合音からの選択的分離抽出

最後に、実際の音楽音響信号に近い状況として、四種類の楽器音が混合した信号から、特定の楽器音を選択的に分離抽出するシミュレーションを行った。

分離処理の一例を図 6 に示す。混合した信号は、フルート(A4)、ヴァイオリン(C4)、ピアノ(B3)、ホルン(E♭2)の 4 種類である。目的音はフルートとした。この例では入力信号の SNR は 0 dB である。

分離抽出の結果、出力信号の SNR は 8.38 dB となった。これに対しトップダウン処理のみでの SNR が 5.73 dB であるので、提案法の分離処理によって約 2.65 dB の改善効果があったことになる。また、このような

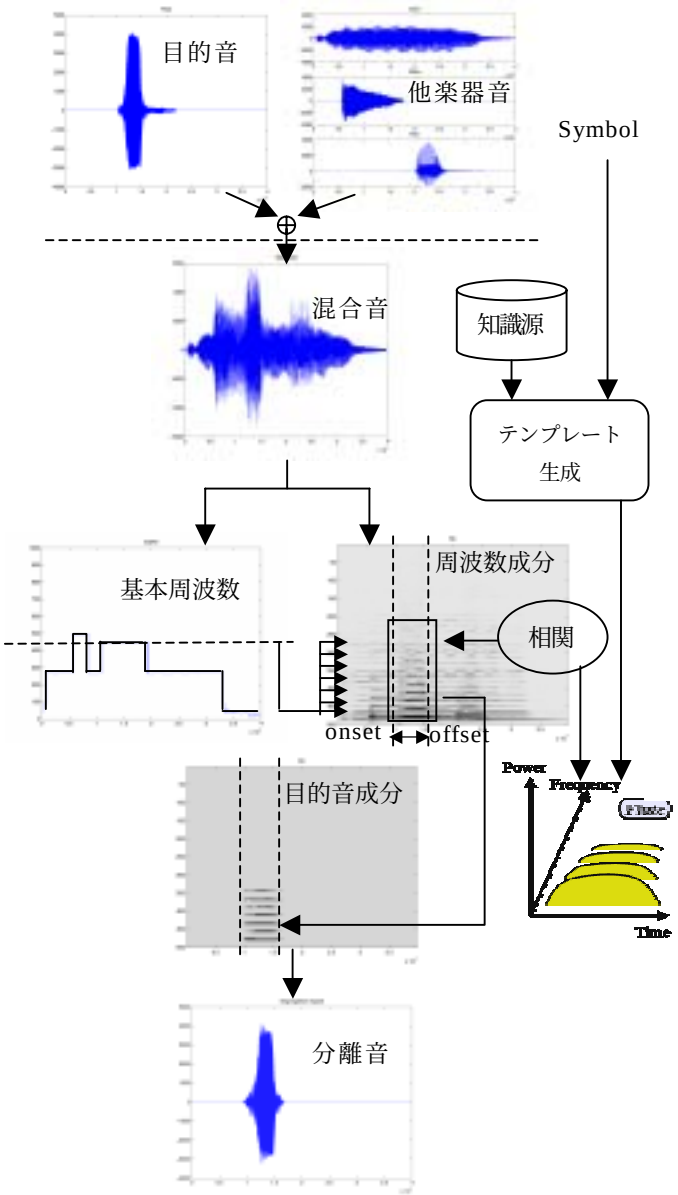

図 6　複数楽器音からの選択的分離抽出の例

混合音に対してボトムアップ処理のみでは目的音を特定できず正しく抽出することができない。それに対し提案法では、複数音源が混合された信号からでも、目的音を選択的に分離抽出できることを確認できた。しかし、フルート以外の楽器音を目的音とした場合、条件によっては基本周波数を正しく推定できない場合があった。これについては、今後の課題とする。

5. まとめ

本論文では、“聞耳”を表現する一手法として、目的音の音響的特徴を知識として利用した選択的な音分離抽出法を提案した。これは、テンプレートを利用した目的楽器の選択を行うトップダウン処理と、他音源成分との分離を行うボトムアップ処理を融合することで選択的分離抽出モデルを実現した。

提案法の有効性を示すために、目的音と白色雑音の混合信号と二楽器音混合信号について分離抽出シミュレーションを行った。その結果、トップダウン処理、ボトムアップ処理単体での結果と比較し、分離精度の向上が確認できた。また、四種類の楽器音が混合した信号からの分離シミュレーションを行った結果、複数の楽器音からの選択的な分離抽出が可能であった。以上より、選択的な分離抽出を行う方法として、目的音の特徴量テンプレートを利用する本手法が有効であることを示すことができた。

今後は、(1) 複数の楽器混合音からでも目的の楽器音を正確に選択、分離抽出するための基本周波数推定法の改良、(2) 分離対象を現在の楽器音単音のみから連続した楽器演奏音へ拡張するために、基本周波数推定部への音階やメロディといった知識情報の適用、(3) 最適なテンプレート作成法の確立、などに取り組む予定である。

謝辞　本研究の一部は、文部科学省科学研究費補助金(No.14780267)の援助を受けて行われた。

文　献

[1] 赤木正人, “カクテルパーティ効果とそのモデル化,” 電子情報通信学会誌解説, Vol.78, No.5 pp.450-453, 1995.

[2] Ellis, D. P. W. , Prediction-driven computational auditory scene analysis. Ph.D. thesis, MIT Media Lab, Massachusetts, 1996.

[3] 中谷智広, 川端 豪, 奥野 博, “計算論的アプローチによる音響ストリームの分離,” 音響学聴覚研究資, H-93-83, Dec. 1993.

[4] 鵜木祐史, 赤木正人,“聴覚の情景解析に基づいた雑音下の調波複合音の一抽出法,” 電子情報通信学会論文誌, Vol.J82-A, No.10, pp.1497-1507, 1999.

[5] Bregman, A.S., “Auditory Scene Analysis: hearing in complex environments,” in Thinking in Sounds, ed. S.McAdams and E.Bigand, pp.10-36, Oxford University Press, New York, 1993.

[6] 北原鉄朗, 後藤真孝, 奥野博, “音高による音色変化に着目した音源同定手法,” 情報処理学会 音楽情報科学研究会 研究報告, 2001-MUS-40-2, Vol.2001, No.45, pp.7-14, May 2001.

[7] Unoki, M. and Akagi, M. “Signal Extraction from Noisy signal based on Auditory Scene Analysis,” In Proc. ICSLP98, Dec 1998.

[8] 柏野邦夫, 中臺一博, 木下智義, 田中英彦, “音楽情景解析の処理モデル OPTIMA における単音の認識,”電子情報通信学会論文誌, Vol.j79-D, No.11, pp.1751-1761, 1996.

[9] 木下智義, 坂井修一, 田中英彦, “周波数成分の重なり適応処理を用いた複数楽器音の音源同定処理,”電子情報通信学会論文誌, Vol.J82-D-Ⅱ, No.4, pp.1073-1081, 2000.