音声情報の能動的再構築処理に関する研究 : 騒音環境における音声認識を支援するシステム

勝瀬, 郁代

https://doi.org/10.15017/1398257

出版情報：九州芸術工科大学, 2001, 博士（工学）, 課程博士
バージョン：
権利関係：

Chapter3

動的音脈形成モデルと能動的聴覚体制化の工学的実現

本章では、第2章で得られた新しい心理物理学的知見に基づいて、スペクトルの追跡・予測モデルを新たに構築する。さらにこのモデルを系列的統合過程へ導入し、計算機による原初的聴覚情景分析を実現する。

本研究で提案する計算機による原初的聴覚情景分析は、調波性の手がかりを用いた周波数統合過程とスペクトル追跡・予測モデルを導入した系列的統合過程から構成され、周波数統合過程の結果の信頼性に基づいて列的統合過程を動的に制御することにより、能動的な聴覚体制化を実現する。

3.1 システムの概要

図3.1に提案システムの概要を示す。初めに混合音を時間－周波数分析し、複数の音の基本的要素(音のオブジェクト)へ分解する。周波数統合過程では、調波性の手がかりに基づいて周波数成分をグルーピングし、各時刻でオブジェクトを再構築する。周波数統合過程における周波数成分のグルーピングは常に成功するとは限らない。そこで本研究では、周波数統合過程の信頼性の尺度を導入する。次に系列的統合過程において、周波数統合過程により得られた音のオブジェクトを時間的に統合する。系列的統合過程として、スペクトル追跡・予測モデルを導入する。その結果、系列統合過程の出力は、観測スペクトルと予測スペクトルの重み和として生成される。この重み値を周波数統合過程の信頼性によって動的に制御することにより、能動的な聴覚体制化を実現する。

以下の各節において、周波数統合過程、系列的統合過程の具体的な実装方法について説明する。

なお、本章で述べるシステムは、対象とする混合音は２つの音源から生成された音から構成されると仮定している。しかし、３つ以上の音源がある場合にも適用できるように、システムを簡単に拡張することができる。

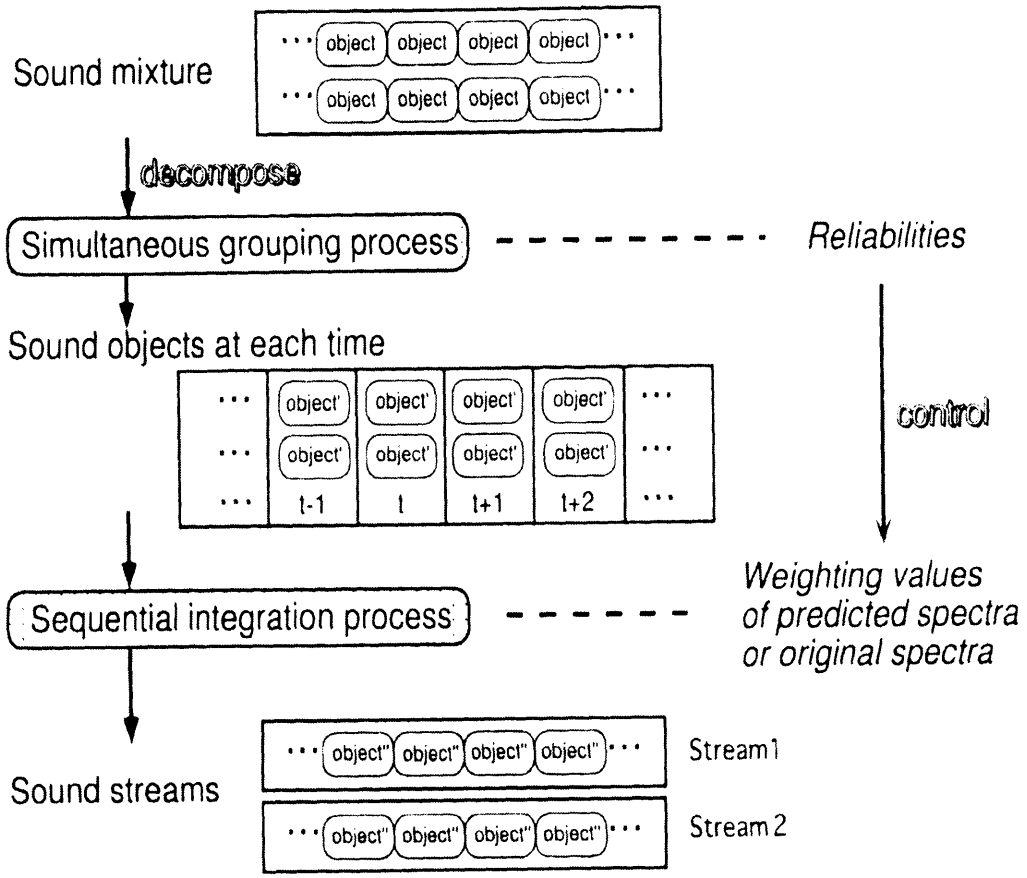

図 3.1: 3章で提案するシステムの概念図

3.2 周波数統合過程の実装

周波数統合過程は調波性の手がかりを用いて実現される。実装面ではピッチ抽出処理と櫛形フィルタによって構成される。

3.2.1 ピッチ抽出処理

ピッチ推定にはTEMPO[64][65]を用いた。TEMPOでは、基本周波数は最小のFMとAMを持つフィルタの出力の瞬時周波数として定義され、基本周波数だけでなくその「基本波らしさ」も同時に求められる。この「基本波らしさ」は、基本周波数の推定誤差の2乗平均の対数に比例する。入力信号が混合音の場合は、構成する音源の基本周波数のうち最も低い周波数が求められる。

3.2.2 櫛形フィルタを用いた調波構造のキャンセレーション

櫛形フィルタを用いて調波性の手がかりに基づくスペクトル成分の分離を行う。櫛形フィルタを用いてスペクトル成分を分離するための方法は2種類ある。一つは強調(enhancement)であり、基本周波数の倍音成分を強調するものである。もう一つはキャンセレーション(cancellation)であり、基本周波数の倍音成分をキャンセルするものである。本システムでは、心理学的知見[66][67]に基づき、キャンセレーションを採用する。図3.2に、キャンセレーションに基づく周波数成分の分離の概略を示す。

混合音が2つの音源の混合から成る時、キャンセレーション処理は混合音を構成する音源の一つの基本周波数を基準とした倍音周波数成分をキャンセルすることによって他方の音源を抽出する。それゆえ、基本周波数の抽出の失敗はそのまま分離の失敗へつながる。TEMPOにおいて、櫛形フィルタから漏れたキャンセレーションの残余の2乗平均値は基本周波数の相対的な推定誤差の2乗平均値におおよそ比例する。ゆえに、周波数統合過程の信頼性と基本周波数らしさの関係は次式のようになる。

$$reliability \propto 10^{-fundamentalness} \tag{3.1}$$

周波数統合過程では、観測信号に対してそのまま櫛形フィルタを適用するのではなく、ウェーブレット変換を用いて入力信号をサブバンド信号$s(t,f)$に分割した後、それぞれのサブバンド信号に対して適用する。ここで、tは時刻、fはサブバンドの中心周波数である。

時刻tでの基本周波数を$f_0(t)$とするとき、サブバンド信号に対する櫛形フィルタの出力$\hat{s}(t,f)$は次式のように与えられる。

$$\hat{s}(t,f) = s(t - \frac{1}{2f_0(t)}, f) - s(t + \frac{1}{2f_0(t)}, f) \tag{3.2}$$

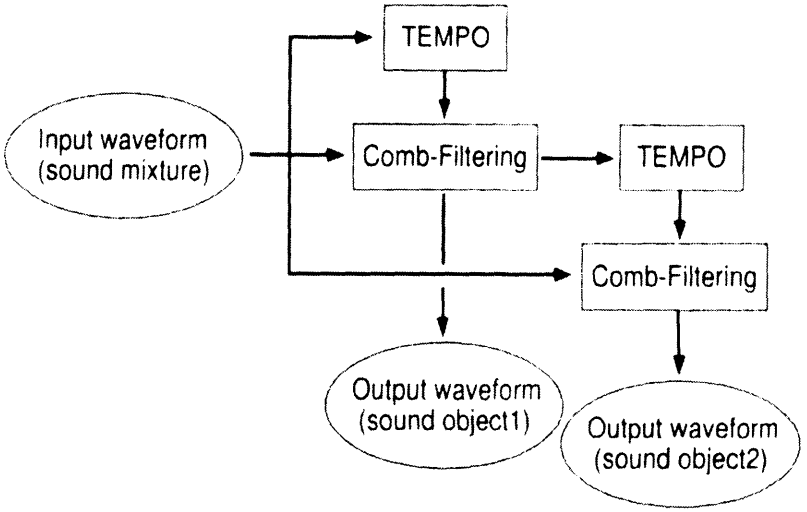

図 3.2: 周波数統合の実装

そして、出力$\hat{s}(t)$は次式のように与えられる。

$$\hat{s}(t) = \frac{1}{C}\sum_{f}\hat{s}(t,f) \tag{3.3}$$

ここでCは正規化定数である。

3.3 系列的グルーピング過程の実装

系列的グルーピング過程では、第2章の心理物理学的実験で示唆された動的なスペクトル追跡・予測過程を導入する。

これまで、系列的統合の手がかりを用いることで、周波数分割されたフィルタの出力を個々に追跡する方法がいくつか提案されてきた[46][47][45]。これらの系列的統合の方法はチャンネル間アプローチと呼ばれるものである。しかし、このアプローチにおける周波数帯域の選択には不確定性が残されている。周波数帯域フィルタの帯域幅が大きい場合は、音声のスペクトル的な特徴がぼやけてしまう。また帯域幅が小さい場合は、音声のような調波構造を持つ信号では、倍音構造と声道伝達特性の相互作用によって帯域間のエネルギー遷移が不連続的になり、連続性の評価の信頼性は下がる。

また従来、音声を記述する特徴としてホルマント周波数が当然のように用いられてきた。しかしながら、その聴覚的実体は依然つかめておらず、むしろ生理学的な観点からは、聴覚情報表現からホルマント周波数を抽出することは困難であることが指摘されている[68]。また、音声の情

報表現としてスペクトル包絡を適切に保持することは、第4章で実装する音声知識との照合過程でも必要となる。そこで本研究では、追跡・予測する対象として、Aikawaら[53]のようにホルマントという特異な周波数のみに焦点を当てるのではなく、スペクトル包絡を追跡・予測する。そして、そのスペクトル包絡を、音源の周期性の影響の除去と分解能の確保を両立させる手法を用いて求めることにより、音声の周波数的特徴を失うことなく、連続性の評価を可能にする。

本研究で提案するスペクトル追跡・予測モデルは2つの過程から構成される。一つは、周波数統合過程の出力をどの音脈に割り当てるかを決めるための割り当て過程(allocation process)である。もう一つはスペクトルを追跡・予測する過程である。

3.3.1 音脈の開始並びに終了点の同定と音脈の数の同定

追跡の開始と終了はパワーではなく、基本周波数の「基本波らしさ」に基づいて決められた。音脈の数は高いレベルを持つ基本波らしさの数によって決められた。

3.3.2 音脈への割り当て処理の実装

周波数統合過程で得られた出力オブジェクトを音脈に割り当てる方法を決めるうえで、参考となる心理物理的知見がある。DivenyiとAlgaziは基本周波数とスペクトル重みが異なる2つの音の知覚的結合程度を調べた[69]。その結果、基本周波数の距離とスペクトル重み距離はトレードオフの関係にあった。このトレードオフの性質については現在のところまだよくわかっていない。本研究では、最も単純な距離尺度を用いてこのトレードオフを表現する。具体的には、周波数統合過程の2つの隣接する時間の出力の任意の2つの音の組み合わせのケプストラム距離とピッチの差の積を求め、その差が最小となる組み合わせによって、音脈への割り当てを行った。

3.3.3 スペクトル包絡の抽出

スペクトル包絡はSTRAIGHT [65]を用いて求めた。STRAIGHTでは、時間分解能と周波数分解能の積が最小で、かつそれぞれの比が等しくなるような時間窓を用いて分析したスペクトログラムに対して、双一次曲面を保存し、かつ時間−周波数方向での広がりが最小となる補間関数を用いた補間操作を行うことで、音源の周期性の影響の除去と分解能の確保を両立させている。

具体的には、窓の時間長が基本周期に応じて適応的に変化する窓関数

$$W(\omega) = \frac{\tau_0}{\sqrt{2\pi}} exp(-\pi(\omega/\omega_0)^2) \tag{3.4}$$

を用いて分析されたスペクトログラム$F(\omega, t)$に対して、次式で定義される補間操作を行う。

$$S(\omega, t) = \sqrt{g^{-1}(\int\int h_t(\lambda, \tau) g(|F(\omega - \lambda, t - \tau)|^2) d\lambda d\tau)} \tag{3.5}$$

$g()$ は単調な連続関数で、補間操作を通じて保存される量を決めるために用いられる。例えば、$g()$ として0.3乗則を用いれば、信号のラウドネスを近似的に保存することができる。補間関数 $h_t(\lambda, \tau)$ は、時間方向の長さが信号の基本周期の２倍であるようなBartlett関数と、周波数方向の長さが基本周波数の２倍であるような三角形の関数の積として表される。

3.3.4 スペクトルの変化の追跡と予測 (スペクトル追跡モデル)

スペクトル包絡はケプストラムへ変換される。STRAIGHTを用いて得られたスペクトル包絡は４０次程度のケプストラム係数で十分記述できるため[70]、最初の４０のケプストラム係数のみを用いる。それぞれの次数のケプストラム係数は時系列とみなされ、時系列解析手法により解析される。ここでこの時系列がどのようなモデルに従うものとするのかを決めなければならない。前章で述べたように、Aikawaらによって報告されたピッチ追跡モデルは２次のAR特性を持っていた[53]。また、赤木らは[54]、２次の臨界制動モデルを用いてスペクトルの変化のターゲットを予測することにより、母音調音の過小評価 (underestimation) の回復を試みた。これらの研究から、本研究においても時系列は２次のARモデルに従うと仮定することが適当であると思われる。

状態はカルマンフィルタを用いて推定され、時系列は予測、補間される。時系列信号を y_n とするとき、状態空間モデルは次の２式を用いて表現できる。

$$x_n = F_n x_{n-1} + G_n v_n \tag{3.6}$$

$$y_n = H_n x_n + w_n \tag{3.7}$$

(3.6) 式はシステムモデルと呼ばれるもので、x_n は直接観測できない状態、v_n はシステムノイズである。(3.7) 式は観測モデルと呼ばれるもので、w_n は観測ノイズである。x_n がARモデルに従うとき、

$$x_n = F x_{n-1} + G v_n \tag{3.8}$$

であり、

$$F = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \\ 1 & & & \\ & \ddots & & \\ & & 1 & 0 \end{bmatrix} \tag{3.9}$$

また、

$$G = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.10}$$

となる。状態x_nの最初の要素は観測信号y_nに相当するので

$$H = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}. \tag{3.11}$$

カルマンフィルタでは、状態x_nの条件付周辺分布が次式のように再帰的に計算される。

$$x_{n|n} = K_n y_n + (I - K_n H_n) x_{n|n-1} \tag{3.12}$$

ここでK_nはカルマンゲインである。条件付き分散共分散行列がV_{nj}、観測ノイズw_nの分散がR_nであるとき、カルマンゲインK_nは、

$$K_n = V_{n|n-1} H_n^t (H_n V_{n|n-1} H_n^t + R_n)^{-1} \tag{3.13}$$

となる。

提案システムでは、観測ノイズw_nの分散R_nが周波数統合過程の結果の信頼性によって決定される。それゆえ、(3.12) 式は時刻tでの出力スペクトルが2つのスペクトルの重み和として与えられることを示している。一つは周波数統合過程の出力であり、もう一つは、時刻$t-1$までに得られたデータを用いて推定された状態x_nによって予測されたスペクトルである。さらに、(3.13) 式は重み値が周波数統合過程の結果の信頼性によって動的に制御されることを示している。提案システムでは、状態は現時点までに音脈に割り当てられたすべての入力を用いて再推定される。それゆえ、推定の精度は時間とともに向上する。

3.4 実装例

本節では、提案モデルを用いて2種類の知覚現象を模擬する。一つは音韻修復であり、もう一つは同時発話音声の分離である。

3.4.1 音韻修復

ここでは原初的聴覚情景分析による音韻修復の模擬を行う。

男声発話音声/eiyu/の一部を50msの白色雑音に置換したものをシステムの入力とした。雑音置換部では基本周波数の基本波らしさは大きく低下するので、スペクトルの追跡・予測過程では、出力スペクトルにおける予測スペクトルの割合が大きくなる。

図3.3は入力信号のスペクトログラムと波形を、図3.4は雑音置換前の元音声/eiyu/のスペクトログラムと波形を示す。図3.5は雑音置換音声を入力としたときの、提案システムの出力音のスペクトログラムと波形である。

原音声とSTRAIGHTを用いて再合成された音を聞き比べたところ、合成音声は元音声と知覚的に区別できなかった。提案システムの顕著な特徴は、「失われた音」は元の状態へ修復されるの

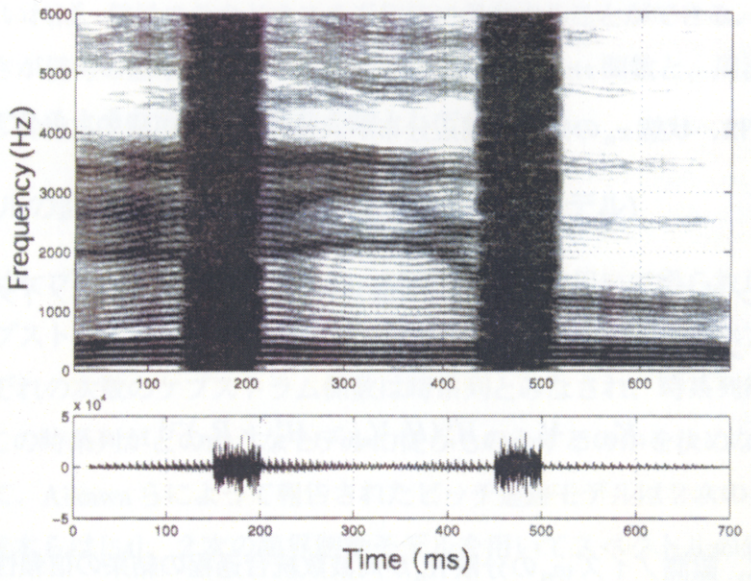

図 3.3: 入力音声
一部が50ms のホワイトノイズで置換された
男声発話/eiyu/のスペクトログラムと波形。

ではなく、知覚的連続性を維持するように修復されることである。修復された音は物理的に元音声と同じである必要はないので、スペクトル歪といった物理量を用いた評価方法は、この場合には適切ではないであろう。

3.4.2 2話者音声の分離

提案システムを用いて音声分離の模擬を行った。システムへの入力信号は男声発話/eiyu/と女声発話/eiyu/の混合音である。

混合音と元音声のスペクトログラムと波形を図3.6, 3.7, 3.8 に示す。原音声の基本周波数、並びに混合音声から求められた基本周波数と基本波らしさを図3.9, 3.10 に示す。

図3.11 と3.12 は櫛形フィルタの出力である。図中 (a) で示された領域は音脈の開始や終了に起因する分離の失敗を示し、(b) で示された領域は2つの基本周波数が倍音関係であることに起因する分離の失敗個所である。図3.10 を見ると、(a)(b) に相当する時刻の周辺で基本波らしさが低下していることがわかる。

図3.13 と図3.14 は提案システムを用いて分離した結果を示している。図3.11 や3.12 でみられた周波数統合過程で分離に失敗した個所は、その位置を特定できないほど修復されていることがわかる。

出力スペクトログラムからSTRAIGHT を用いて再合成された音を聴取したところ、原音声と

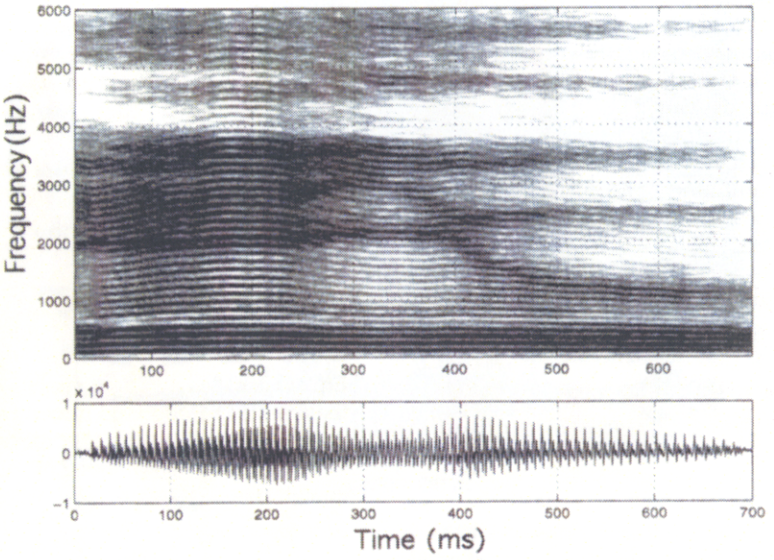

図 3.4: 男声発話/eiyu/のスペクトログラムと波形

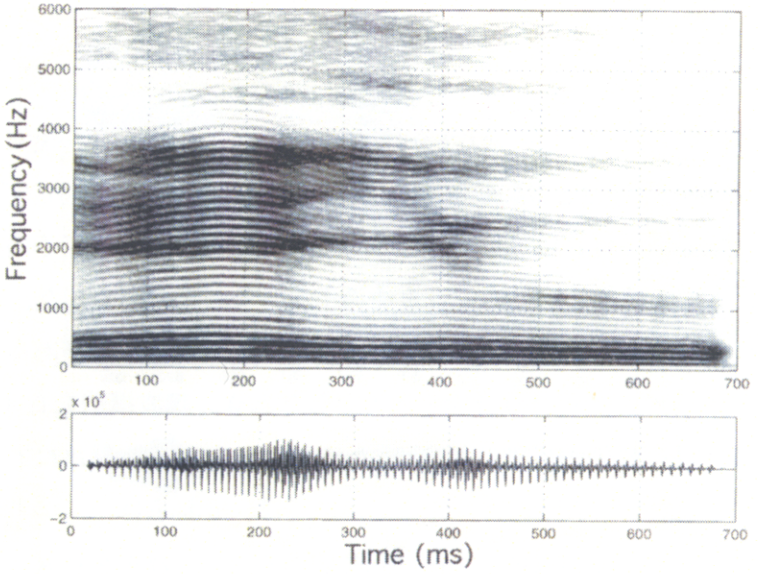

図 3.5: 出力音のスペクトログラムと波形

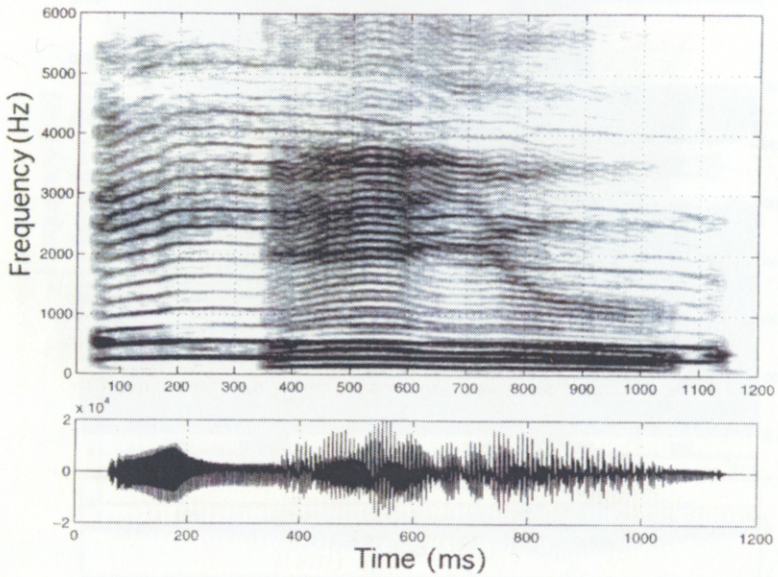

図 3.6: 入力された混合音のスペクトログラムと波形
男声発話/eiyu/と女声発話/eiyu/の混合である。

比較してわずかに音質の劣化が見られた。これは、音声には非周期成分が本質的に含まれているため、キャンセレーション操作では十分にキャンセルされずに残余している成分があったためである。しかし、再合成音は十分に発話内容を同定することが可能な音質を維持していた。

3.5 まとめ

本章では、第2章で示唆された聴覚の新しい知見に基づいた計算モデルを提案した。これを原初的聴覚情景分析の系列的統合過程に導入することにより、周波数統合の主要な手がかりである調波性の手がかりによる分離に失敗した場合でも、系列的統合過程によってその欠落した部分を能動的に補うことによって、知覚的に妥当な音声を再構築することに成功した。

原初的聴覚情景分析の計算モデルのみで音声の分離抽出を行うには限界がある。原初的聴覚情景分析の法則では、例えば／sa／という音節を入力としたとき、調波構造を有する／a／といった音と、ノイズバースト性の音である／s／といった音は別の音脈へ分凝されてしまい、音声として一つの音脈が形成されない。人間の音声知覚においては、それらは一つの音脈として知覚されることから、音声スキーマの強い貢献の存在が示唆される[15]。音声を対象とした音脈形成においては、音声スキーマの貢献は大きいとしても、音声の持つ特徴量に対して基礎的な群化と分凝を成し遂げるのはやはり原初的聴覚情景分析過程であり、本システムはその過程において、さらに能動的な体制化を実現する枠組みを導入したという点で、意義があるものと考える。

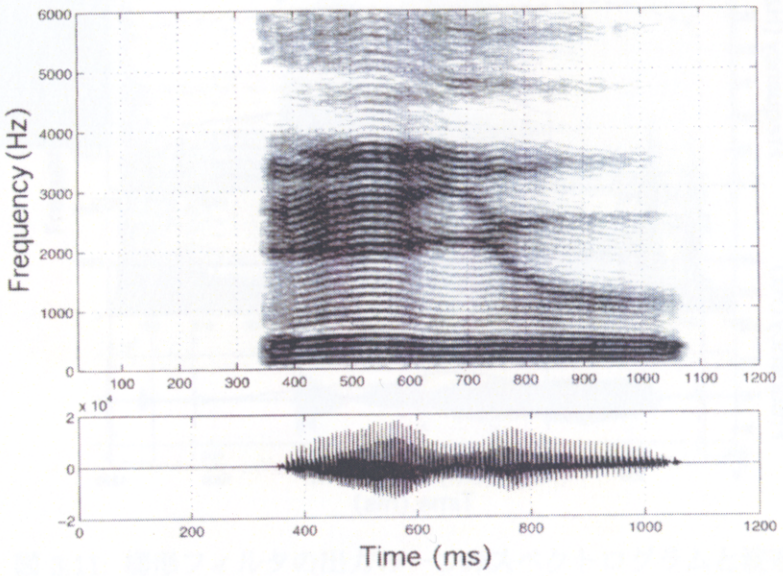

図 3.7: 混合音に含まれる男声発話/eiyu/のスペクトログラムと波形

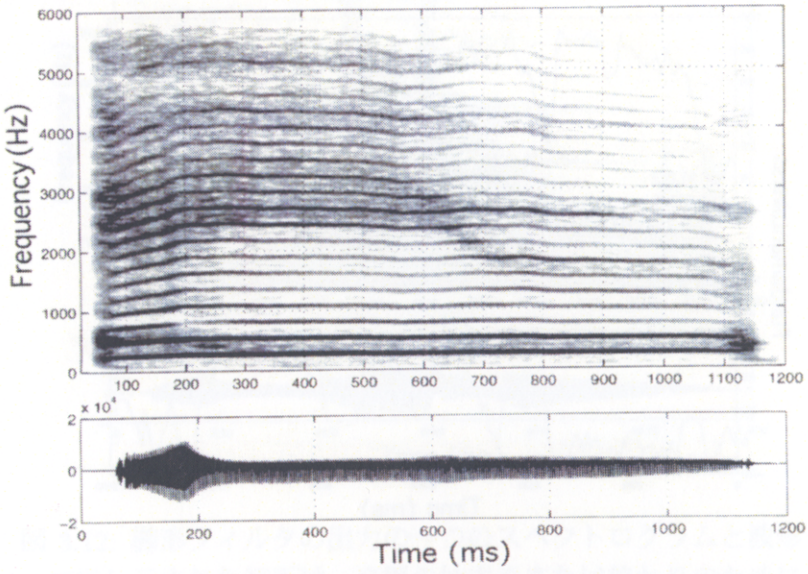

図 3.8: 混合音に含まれる女声発話/eiyu/のスペクトログラムと波形

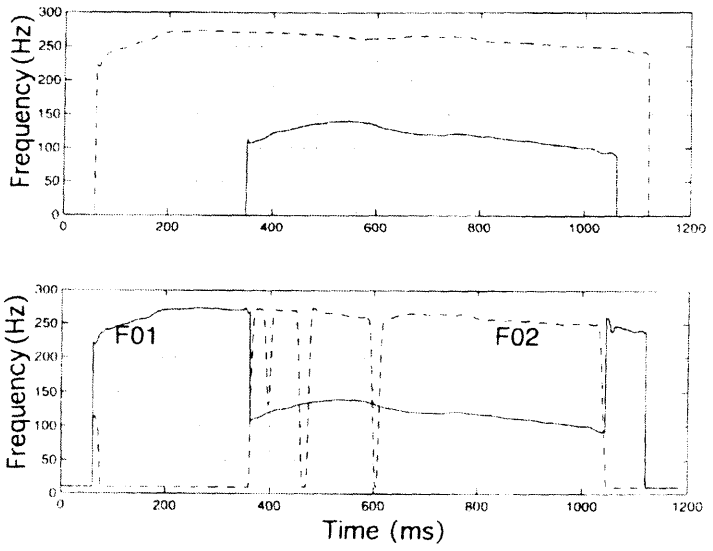

図 3.9: 原音声の周波数(上図)と混合音から抽出された基本周波数(下図)
上図では、実線が男声/eiyu/の基本周波数で、破線が女/eiyu/の基本周波数である。下図では、最初に求められる基本周波数(F01)を実線で、次に求められる基本周波数(F02)を破線で示す。

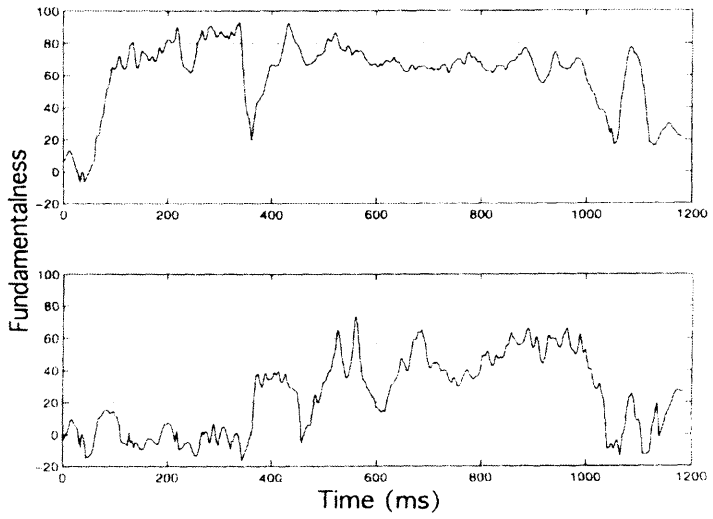

図 3.10: 求められた基本周波数の「基本波らしさ」。
上図は図3－9のF01の基本波らしさであり、下図はF02の基本波らしさである。

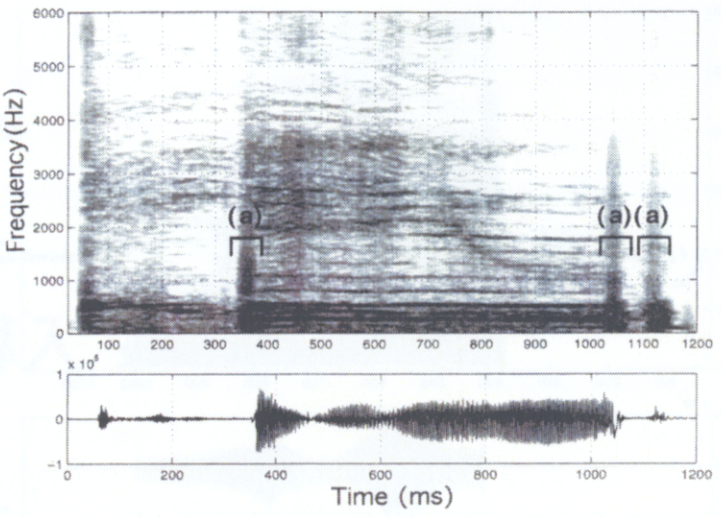

図 3.11: 櫛形フィルタの出力の一つのスペクトログラムと波形
(a) で指し示された個所は、音脈の始まりまたは終わりのために、
分凝に失敗している個所である。

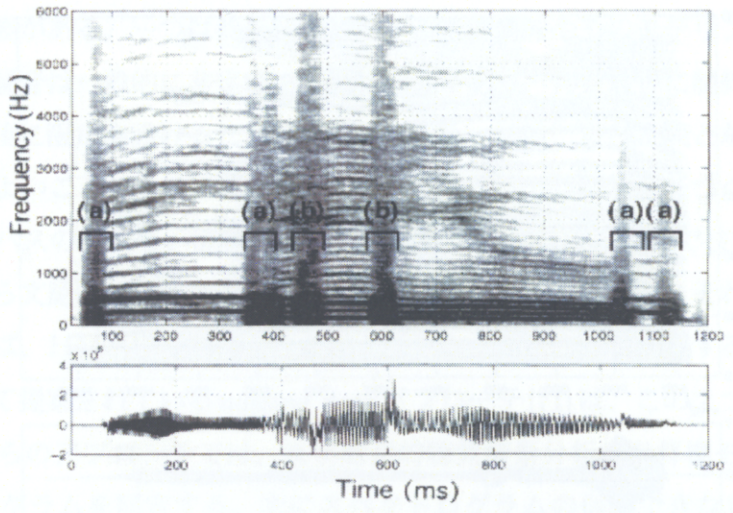

図 3.12: 櫛形フィルタの出力の一つのスペクトログラムと波形
(a) で指し示された個所は、音脈の始まりまたは終わりのために、
分凝に失敗している個所である。(b) で指し示された個所は、基本
周波数が倍音関係になったために分凝に失敗している個所である。

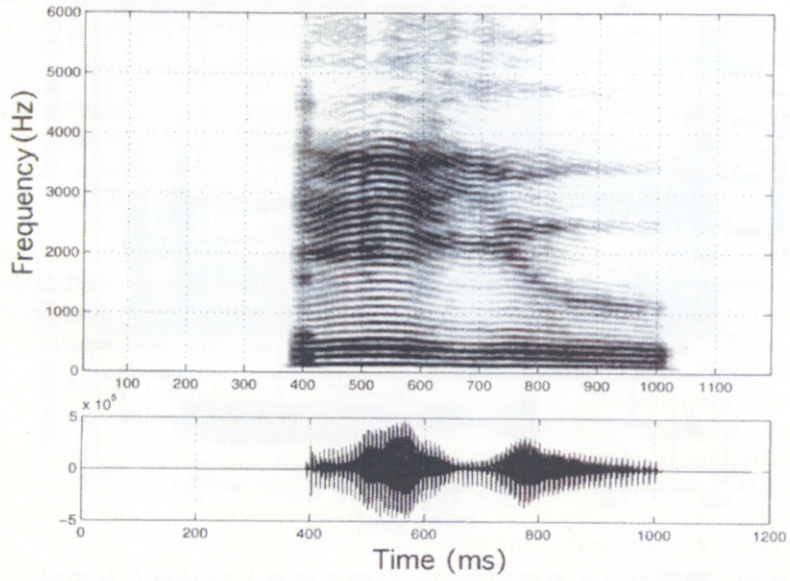

図 3.13: 出力音の一つのスペクトログラムと波形
混合音を構成する音声のうち、男声発話/eiyu/に相当する。

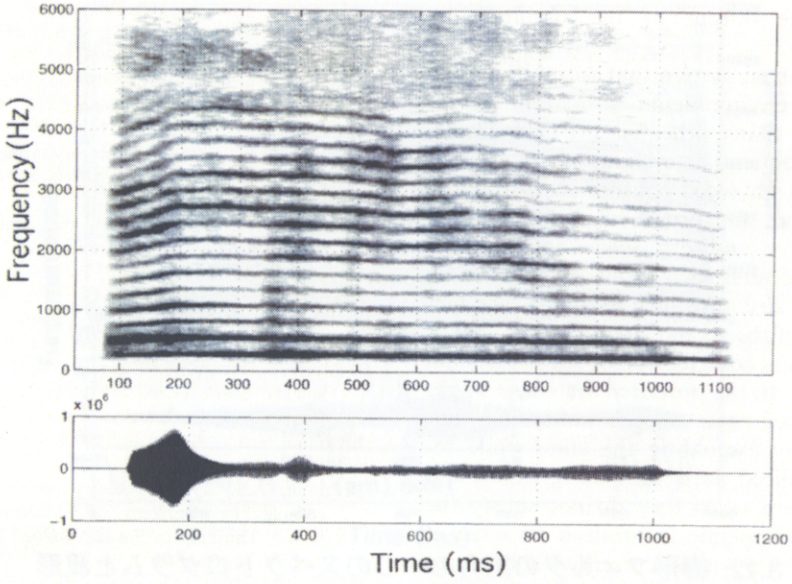

図 3.14: 出力音の一つのスペクトログラムと波形
混合音を構成する音声のうち、女声発話/eiyu/に相当する。